MEASURING MINORITY LANGUAGE PRESENCE ON WIKIPEDIA – TOWARDS THE MEASUREMENT OF MINORITY LANGUAGE PRESENCE ON THE WORLD WIDE WEB

ANDREW DEERE

Faculty of Advanced Technology
University of Glamorgan
Pontypridd, Wales, UK

A submission presented in partial fulfilment of the requirements of the University of Glamorgan/Prifysgol Morgannwg for the degree of Doctor of Philosophy

ABSTRACT

Measuring Minority Language Presence on Wikipedia – Towards the Measurement of Minority Language Presence on the World Wide Web.

This study examines twenty majority language and minority language versions of Wikipedia (English, French, German, Spanish, Italian, Danish, Icelandic, Welsh, Gaelic, Scots Gaelic, Manx, Cornish, West Frisian, North Frisian, Saterland Frisian, Breton, Catalan, Galician and Sardinian). The results are used to compare the majority language versions with the minority languages that compete with them. The statistical results are gathered by a custom designed software program that analyses several content metrics (number of words, images and an analysis of the quantity and location of links) relating to each article, allowing for a more accurate measurement of the contents of a language edition of Wikipedia than is usually given by a raw count of the number of articles.

While studying Wikipedia in isolation is an interesting question, there is no difficulty in understanding that the minority language editions of Wikipedia are quite small in both range of articles, as amply demonstrated by a study of Wikipedia's statistics in this regard. The goal is to put forward a potential model for estimating and quantifying minority language material on the web.

Outlined in this study is a definition of 'web presence that is a two-dimensional concept combining both breadth of coverage and depth of content; a formula to measure the presence of a particular language edition of Wikipedia and, by extension, the web; a formula to compare the presence of one language with that of another; a "language constellation" system that measures languages in meaningful groupings based on real world competition model; a "tiered classification" model, that uses presence values to be predictive and descriptive of where a language's presence relates to other languages.

DEDICATION AND ACKNOWLEDGMENTS

I dedicate this work and give special thanks to my mother, Heather, who always believed in me.

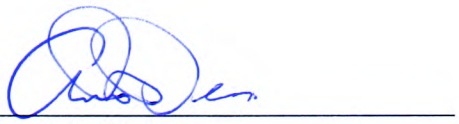
I would like to give especial thanks to my supervisor, Daniel Cunliffe who was always ready to give me much of his time and offered invaluable assistance, support and guidance over the many years of undertaking this study. The errors in this work are entirely the author's and

I would also like to thank Dr. Kelly Holmes who read and provided excellent comments on my Transfer Report.

Many thanks also go to my thesis committee: Diarmait Mac Giolla Chríost and Duncan McPhee who challenged many of the points and whose advice resulted in several important changes to produce this revised version of the original draft.

CERTIFICATE OF RESEARCH

This is to certify that neither this thesis nor any part of it has been presented or is currently being presented in candidature for any other degree than the degree of Doctor of Philosophy of the University of Glamorgan

CANDIDATE 

Andrew Deere

May 2011

CERTIFICATE OF RESEARCH

This is to certify that, except where specific reference is made, the work presented in this thesis is the result of an investigation undertaken by the candidate.

CANDIDATE

Andrew Deere

DIRECTOR OF STUDIES

Daniel Cunliffe

Table of Contents

Figures

Tables

Equations

I. Introduction

This general purpose of this study is to quantify the availability of materials in 'minority languages' (as defined below) on the World Wide Web[1]. Since studying the entirety of the web is difficult, this study concentrates one significant web application – Wikipedia. While this study focuses on Wikipedia, it is not the intention to provide a picture of the multilingual nature of Wikipedia *per se* (although that is a useful result), rather it is hoped that this may provide a model that, if it can be demonstrated to work on one part of the web, could be refined and extended to other areas of the web. This would allow us to study minority language use on the web on a scientific basis.

A. 'Minority Languages'

By some estimates there are up to 7,000 languages spoken throughout the world today (Crystal, 2002; Ethnologue.com, 2008; Grenoble & Whaley, 2006: 1). Some of these languages are thriving and the number of active users is increasing. A dozen or so languages have hundreds of millions of speakers and vast amounts of written, audio and visual material. Another few dozen, perhaps a hundred languages, mostly official languages of European and Asian states, are relatively stable and have enough speakers and enough material to ensure their long term survival. But the vast majority of languages spoken and used today are in a precarious situation: they have either a small, and often decreasing, base of speakers or are subject to a number of forces that conspire to reduce their use and scope. In the opinion of a number of researchers, many languages will become extinct before the end of this or the next century (Abley, 2004; Crystal, 2002; Fishman, 2000; Janse & Tol, 2003; Lewis & Simons, 2009; Ó Néill, 2005).

Various terms have been used to describe the thousands of disadvantaged languages that are in danger of becoming obsolete if present trends continue: 'lesser-used', 'regional', 'heritage', 'endangered' and 'minority' are terms often used in the specialised literature. Some languages are 'lesser-used' in the sense that the number of speakers of those languages

[1] While opinions differ, the words 'web' and 'internet' will be in miniscule initial letters in this document. The web and the internet are not nouns for proprietary products, nor are they acronyms, and normal English usage would not require their capitalisation.

is not large, but the language may be in constant use by a community of speakers and thus relatively stable. A lesser-used language that is not the official language of a state, but is spoken in a definable locality, may be called a 'regional' language. Such regional languages may be quite large in terms of the number of speakers, but may be under some pressure from a competing language or, by not having equal status, the language is otherwise under threat. A 'heritage' language is one that is still spoken, though often in a limited range of activities, oftentimes the former normal language of a community, but now only spoken in limited circumstances. An 'endangered' language is one whose number of speakers is decreasing rapidly and often is not being passed on to the younger generations. A 'minority' language is a language spoken "by a group which is significantly smaller in number than the rest of the population."(Crystal, 2008:307) A minority language is a language used by a minority of people within a given geographic area, mostly in situations where a dominant (i.e. a 'majority' language) is used in the same area. Frequently, when a minority and a majority language are in competition, the minority language suffers a reduction in its spheres of use causing a decline in the number of speakers. This usually, but not necessarily, implies that the minority language is disadvantaged. The term 'minority language' refers not only to the relative number of speakers but rather to the status, or more precisely the lack of status, of the language.

To simplify matters, in this dissertation, the term 'minority languages' will be employed to designate all smaller languages that, to a greater or lesser degree, face difficulties and which could be more precisely described by the other terms mentioned above. This will include all languages that the European Commission labelled "regional/minority languages", as set out in the following Table 1(European Commission, 2006).

Noted linguist David Crystal has written four books that argue that there is a "language revolution" underway that affects all the languages currently spoken and used. In three specific books he discusses the rise of global English in the modern world (Crystal, 1997); the changes that languages will undergo occasioned by the invention and rise of the internet (Crystal, 2001); and the accelerating rate of extinction of a large number of the world's languages (Crystal, 2002). After publication of these three works, Crystal reworked the themes presented in his trilogy into a more general volume entitled *The Language Revolution (Crystal, 2004)*. In Crystal's view, we are witnessing a "revolution" composed of

direct their attention toward saving or promoting the particular language that is of primary concern to them. Oftentimes this is their mother tongue, and they have a focus in protecting or promoting it, leaving to others the same job for their mother tongue. Things are beginning to change, and increasingly, there is an understanding that minority languages often suffer from a similar set of problems, and thus a number of studies are building a solid theoretical basis for minority language studies. Notable in this regard is the pioneering work of Fishman (Fishman, 1991, 2000, 2002) who has provided a set of universal themes that can provide common ground for most minority language researchers.

B. Measuring Languages on the Web

Modern technology will undoubtedly have a part to play in whether any particular language survives the next few centuries. One of the classic attributes of languages that are in danger of disappearing is the relative lack of materials available in that language. This can cause a vicious circle where the use of that language diminishes as the amount of available materials decreases, resulting in a further decrease in the number of speakers, which, in turn, further accelerates the decrease in materials. The web has the potential to affect this dynamic, although, at the moment, it is not clear whether the web will be a positive or a negative. There are two alternatives. With the lowering of the costs associated with publication of written, spoken and visual materials, and its ability to reach wider audiences, the web may provide a wonderful opportunity for minority languages to increase the amount of available material and thereby assist minority languages in their long term survival. Conversely, By enabling the larger languages more scope for dominance, minority language speakers will be ever more overwhelmed by majority language material, leading to acceleration in the decline of some minority languages.

Of course, not all languages are in the same situation and it is possible, and indeed probable, that some languages will find a way to harness the web to their advantage, while others will the fail to meet the challenges. As noted by one researcher in the area: the internet is not neutral, and there are a number of barriers. Some languages will have an easier time; others will not (Paolillo, 2005). As is stated in a UNESCO Institute for Statistics (UIS) report:

> "There is a need to take stock of the current global data with regard to ICT, and to identify any gaps that might exist, in order to help decision-makers within countries draft informed national policies vis-à-vis the Information/Knowledge

> Society. ... how can the world understand its progress without concrete measurements of where we stand currently and without commitment to continue to measure progress? Thus there is an immediate need to put in place reliable data systems and well-defined series of both baseline and repeated data sets and indicators that are capable of giving a quantitative picture of change...(UNESCO Institute for Statistics, 2003: 15-16).

The web is a communication system that links a vast array of public and private computer systems. The range and types of information that can be carried through this system is quite astonishing: written text, images, video, music, databases, computer programs and virtually any other medium that man has created has been successfully transmitted by means of the internet. But it is written text and spoken word that man uses primarily to communicate. To a large extent, the internet is language. More precisely the internet is human language.

Measuring the web is very difficult. Not only is the web a complicated artefact, but it is very large. A number of studies have attempted to show how large the web is (Bergman, 2001; Gulli & Signorini, 2005), although all they could do was provide rough estimates at a particular time. In October 2005, Eric Schmidt of Google, one of the few organisations that had the resources to examine the entirety of the web, estimated the size of the web at that time at 5,000,000 terabytes[2]. At that time, Google had indexed only 170 terabytes. Thus, even if the web has stood still in 2005, Google would have needed another three hundred years for it crawl and index the remaining material.

Some fifteen years after the birth of the web it can be simply stated that the web is very large, is growing daily and has already reached the point that it is impossible to estimate its size with any degree of precision. Furthermore, with the advent of 'Web 2.0'[3], where user generated content and dynamically created pages constitute some of the most popular parts of the web, it becomes difficult to define what it is we are trying to measure. This is in marked contrast to traditional websites and pages, where the user was limited to consuming the content only. Examples of Web 2.0 include social-networking sites, blogs, wikis, video-sharing sites, hosted services and web applications. Further enhancements allow users to

---

[2] http://news.softpedia.com/news/How-Big-Is-the-Internet-10177.shtml

[3] The term "Web 2.0" relates to the title of a conference in 2004 by Tim O'Reilly, the founder of O'Reilly Publishing (see http://tim.oreilly.com/). The term refers to a new way of looking at the Web, chiefly as a result of the maturing of Web design technologies that increasingly made use of dynamically created Web pages

create their own pages (typical of social networking sites such as Facebook, or advanced e-commerce sites such as Amazon.com). Because the user can create their own *ad hoc* pages, or even their own mini websites, it is difficult to know what 'pages', or what content exists at any one time. Web 2.0 renders redundant the naïve notion of estimating the size of the web. The web is not just 'infinite' in the sense that it is so large that it cannot be counted, like grains of sand on a beach. The web is also 'infinite' in the sense that there is no physical or logical dimension to its size or growth. Measuring the 'size' of the web is like measuring the number of words that can be spoken.

Any measurement of the web needs to take the following factors into account: (1) the web is now too big to count, and even if it could be counted at some point in its evolution it has now exceeded that point; (2) the web is no longer composed of discrete objects (webpages, websites) that can be counted, and, even it were so composed; (3) the web is infinite and will continue to grow.

C. The Search for Solutions

The 2005 UNESCO sponsored report '*Measuring Linguistic Diversity*' is the most comprehensive theoretical study on the measurement of language use on the internet (Paolillo, Pimienta, Prado, & al, 2005). The report rejects using crude indicators of language use:

> "We need to move to develop more intelligent indicators. Measuring languages in the overall number of pages on the Web increasingly presents challenges caused by the sheer volume of Web content. (Paolillo et al., 2005: 9)

And further, it generally rejects the principal methodology used up to that date: manipulation of search engine data, although it rejects it for reasons that are different from those that will be discussed in Chapter III – Literature Review, below:

> "We can easily produce a random count of Internet pages by using any number of commercial search engines, but we cannot judge how often Web pages are read or whether the reading of a page helped the reader in any way." (Paolillo et al., 2005: 9).

Three primary problems are identified by the authors:

- standardisation of definitions to achieve international comparability;
- identification of indicators relevant for developed and developing country policies; and

- capacity building at national and international levels to allow quality data to be collected on a regular basis.

The report provides the useful caveat that web pages are the 'supply side' of the web, and that does not mean that they are necessarily ever consumed. It also notes that the web presents some aspects of a free market: websites are developed to meet the needs of a particular audience. If there is little potential audience, websites will not be developed.

The report notes that "English remains the most prevalent language on the Internet, and some very populous languages have little or no representation..." and that "some languages have large amounts of readily accessible digital content. Internet users who speak, read and write such languages have far less difficulty accessing and sharing useful information than speakers of less well-represented languages." (Paolillo, 2005:43)

D. Measuring Language Material on the Web

The difficult task then is to formulate an accurate, reliable and repeatable method to measure language, and specifically minority language, provision on the web. This is a technical question: how to acquire the necessary raw data from the web to allow us to make meaningful measurements. This task is extremely problematic owing to the enormity of the web on the one hand and the often scarce and scattered material produced in any particular minority language. The sheer amount of processing and storage power needed to access this material makes the web a very difficult thing to measure. To further complicate matters, the web has not been designed to be indexed by language. This is a problem that has plagued other researchers who have tried to ask the same question regarding majority language material, and several solutions, none of them entirely satisfactory, have been proposed. These studies form the focus the Literature Review, Part II below.

This study will look at the problem of measuring on-line language materials and will propose a very basic methodology for producing a set of indicators that can tell us how much of a given language is available to users on one part of the web, Wikipedia. This methodology is outlined in Part III - Research Methodology, below and the model for study of the web is proposed in Part V - Theoretical Models. This study does not pretend to be comprehensive, but it does propose a novel approach to measuring minority language on Wikipedia and by extension, to measuring other parts of the web.

The central core of this study of this is to measure the 'presence' of a particular language on Wikipedia; 'presence' being defined as: how much material is available in a particular language, and, how widely spread that material is. 'Presence', as used herein, is a two-dimensional concept: the aim is to measure how much material is available *and* over how many domains that material is present. The central logic is that if a particular language has a range of material, both in breadth and in depth, in all the usual domains of interest of the web then it could be said that it had a 'good' presence. Conversely, if a particular language has some material, albeit significant in terms of depth, in one domain, but very little material in many other domains, or, if a language had a large amount of material across a number of domains, but the content on those webpages was sparse, we could conclude that the language presence would be 'poor'. Thus 'presence' is a defined as a function of the depth or material available in a given language across a broad range of domains of human interest and activity.

In order to concentrate on an area of the web that can be measured, the principal subject of measurement of this dissertation will be Wikipedia. Wikipedia is inherently multilingual and it provides an excellent opportunity to see how the various languages are making use of the web. Wikipedia is also 'equal' in the sense that each language edition of Wikipedia is, for all intents and purposes, the same as any other edition. If there are any differences, it is solely down to the linguistic community that speaks that language and not to any other factors. By using data obtained directly from Wikipedia an attempt will be made to measure the size and scope of Wikipedia's presence for twenty languages, with an aim of addressing some of difficulties involved and also providing some sort of indication of how this part of the web can be measured.

E. Evolution of the Study

The study of minority language provision on-line began, in 2004, with the question of understanding how minority languages were presented on bilingual webpages that contained text in two languages or on bilingual websites. The initial goal was to investigate whether there was any bias in the presentation of the two languages and, if necessary, to suggest ways in which such bias could be eliminated or attenuated. Preliminary work centred on the establishment of a set of guidelines by which minority languages could be given appropriate treatment on either bilingual or multilingual webpages or websites. A body of work was done

in this regard, which result in a conference paper presentation in 2005 (Deere & Cunliffe, 2005) and its subsequent publication in a book chapter in 2009 (Deere & Cunliffe, 2009).

However, it soon became apparent that the problem facing speakers of minority languages, when they entered the online world, was not mainly in the presentation of materials (though much improvement could be made), but in the paucity of materials available in those languages. While there was some minority language material available, the coverage was sporadic, often trivial, and sometimes negligible. Thus, while recognising that minority languages face a number of barriers inherent in the nature of website and webpage design and in the very structure of the web, the research question moved from how to present such material to measuring how much material is actually available.

1. Using Search Engines

Like several other researchers (see Literature Review, Part II), it seemed that there were ready made indexing systems – search engines – that offered an ideal solution for determining how much material was available on the web. Commercial search engines had already performed the difficult task of crawling and indexing very large numbers of webpages, had indexed the results and had made such results available for public consultation. These search engines would return the number of 'results' or 'hits' for a given search term (such 'hits' defined as the number of webpages in the search engine's database containing the search term), and if these results could be harvested and analysed, a cheap and effective data extraction method could be found.

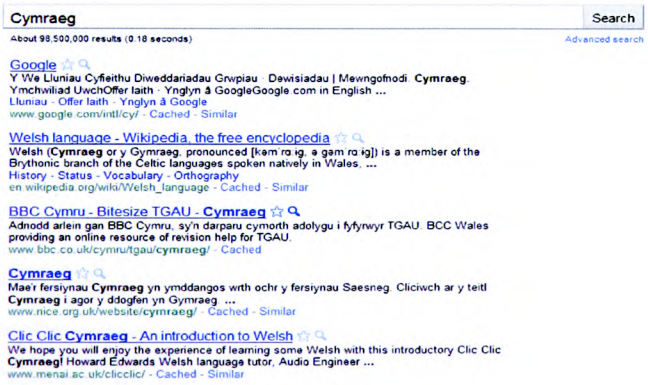


Figure 1 – Sample Google Result Screen (2007)

It was thought that if carefully crafted search terms could be entered into one or more search engines, the number of results could be harvested and analysed, and approximations of the number of webpages in a particular language could be deduced. Using this methodology a program was envisioned that would interrogate the then market leading search engine – Google – and would fetch and store the number of results for a large number of searches. Google had released a SOAP based API[4], which allowed direct, though limited, access to the Google database. Using this API, a program was written, which was able to automate the search for various search terms in one minority language, Welsh, and stored the results for further analysis. By using the program, a search term could be entered, which would then generate a search enquiry that was sent, using the Google API, to the Google database. Google would then return two types of data: a number corresponding to the number of results for the search term, and a list of webpages that contained the search term. The central logic of the program was that if we could obtain from Google the number of returns that its database contained for the same term in different languages, we could then compare the results. For example, if we could compare how many hits contained the English word 'computer' with how many webpages contained the equivalent Welsh word Welsh word - 'cyfrifriadur'- we could estimate the relative presence of English and Welsh on the web.

However, an initial problem was encountered due a grammatical feature of the Welsh language that had a serious effect on the results. In Welsh, as in all Celtic languages, nouns and other parts of speech are affected by 'mutation', which has the effect of changing the spelling of Welsh nouns, without changing the lexical meaning of the word. The Welsh word 'cyfrifiadur' ('computer') can also be spelled 'gyfrifiadur', 'chyfrifiadur' or 'nghyfrifiadur', depending on various rules of Welsh grammar, but in all cases it represents the same lexeme. English has a similar problem in the variant spelling of many English words, usually a British and American spelling: 'computerisation/computerization', for example. But whereas in English the problem is encountered in only a limited number of words, mutation affects a significant proportion of Welsh nouns and other words, including all words that start with vowels and that begin with nine specific consonants. The effect of not taking into account mutation would have the possibility of not counting pages where a mutated form could occur,

[4] http://code.google.com/apis/websearch/, depreciated since November 2010)

but not the radical, unmutated form. Considerable time was therefore taken to ensure that certain the program could adequately deal with this phenomenon.

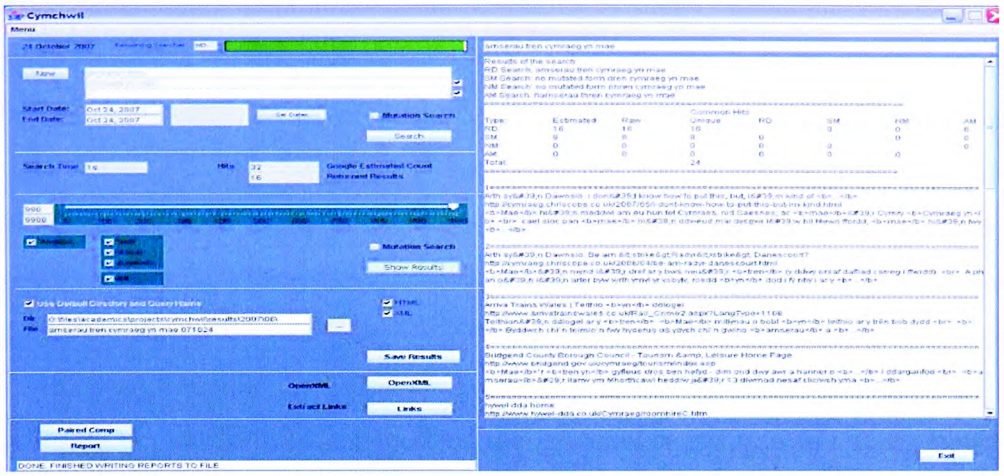


Figure 2 - Screenshot of the Cymchwil Program

Initial tests of the program on search terms that were known to produce small numbers of hits proved quite successful. However, once the initial testing was complete, two different problems were encountered: one technical and one theoretical. The technical problem lay in significant discrepancy between the number of hits reported by Google and the number of webpages returned by Google. For search terms where the number of returns reported was less than 1000 manual verification of the returned list of webpages verified that the Google API was correctly reporting the number of webpages in is database. The margin of error, for terms having less than 1000 returns, was noted to be in the order ±5%. However, when search terms that produced numbers of returns larger than 1000 were attempted, the reported number of returns did not correspond at all with the actual results returned by the Google API. In some cases, Google was reporting several thousand returns, but only returned several hundred webpages. The reason for these discrepancies was never properly ascertained, but lay either in an error in Google's calculation algorithms or in limitations in the Google API.

The second, more theoretical, problem related to the selection of the search terms. Much work went into creating a list of search terms which would enable a search across a broad range of fields of human activity. This list proved to be difficult, in that cultural activities are not uniform across cultures. While it was fairly simple to use generic terms to cover major activities, it proved to be more difficult once one tried to be more specific. For example, what

is the cultural equivalent of a particular city? If one culture tends to follow a particular sport, say football, is it acceptable to substitute another sport, say rugby? However, given the above mentioned technical problem, this index idea was never actually tested, since results could never be confirmed.

At this time that the limitations of using search engines were first encountered and it seemed best at this stage to abandon further research in this area. It was realised that even if the results obtained from Google were accurate, in the sense that the results faithfully represented what data Google had in its database, without knowing how the search engine produced its results, in other words, without knowledge of the algorithms and contents of the dataset, it was impossible to rely on the results. While it was possible to obtain and analyse results from search engines, it could not be know whether those results accurately reflected the actual situation on the web. As work on this problem progressed, the study of literature in the field indicated that others were facing similar and additional problems using the commercial search engine approach.

2. Building a Search Engine

At this point it seemed that the use of commercial search engines was not a viable option. For a period of about six months, a different 'search engine' approach was attempted. Since the principal problem resided in the lack of transparency of the algorithms used by commercial search engines, this problem could be solved if a search engine could be programmed in-house; the program's methodology, algorithms and data set could be known and tested. To do so from scratch would have been an enormous undertaking, but there existed, at the time, an open source search engine suite - Nutch[5] - that could be modified to produce a scrape of a particular webspace and to then analyse this dataset for language indicators. At the time, Nutch was still in development and an early beta version was used (Nutch v. 0.7[6]). While it could be run from the command line, it proved to be difficult to get it to work in a number of development environments for eventual modification. Eventually,

[5] http://nutch.apache.org/

[6] The current version, as of November 2010 is 1.2

after several frustrating months, and with serious doubts as to whether the resources and time were available to allow for proper development, it was decided to press ahead with other forms of analysis. It is believed that this methodology is still sound, and given that Nutch has been developed further in the intervening years, it is possible that further research in this area could yield some usable results.

At this juncture a third program was developed to attempt to measure static websites. This program allowed for the traversal of a bilingual Welsh-English website to an arbitrary level, and, by means of custom built parsers (for Welsh and English only), the program can determine the amount of Welsh and English contained on that site.

```
INFO - *** [1] Crawl: http://www.glam.ac.uk/ Depth: 11
INFO - Finished crawling 35 pages containing 13285 words
INFO - Average number of words in a document: 379
INFO - Timing close index: 1616 ms
INFO - Total Pages crawled: 35
INFO - External links (not crawled): 35
INFO - Welsh pages : 2
INFO - English pages : 33
```

Figure 3 - Log results of a Cymscrape test scrape (2008)

The intent of this program was to analyse how much of a given language was available on a given website. For example, if the website of a governmental organisation that offered a bilingual service could be examined, a comparison of the two language versions could be made in terms of raw content and some conclusions could be drawn about bilingual provision on targeted websites.

However, it became quickly apparent that such a study would be too small in scope. While it might be able to highlight a given organisation's commitment to a minority language, it could hardly tell us how much minority language material there was on the web. Even to the extent that hundreds or perhaps even thousands of websites were studied, we would not be in any position to make generalisations about Welsh language presence. Furthermore, since a selection of appropriate websites would have to be made beforehand, the self-selection problem would affect the statistical analysis. Unless this program were to be used in connection with a search engine that could draw a random selection from the entirety of the web, there would be considerable self selecting in the samples drawn and this would significantly affect the results obtained.

This program performed well at an early stage, but development had to be curtailed due to lack of time and resources. Besides the above mentioned problem, this program suffered from the additional problem that each language would require its own hard-coded parser. This would have lead to significant additional development costs. This program has not used in the preparation of this study, but further work seems promising.

3. Narrowing and Focusing the Enquiry

The study of how to measure minority language use on the web went through a lengthy intellectual trajectory based on the results of practical work and analysis of the literature and through the programming and testing of several methodologies. Since the earlier work together with state of the literature in the field, lead to a conclusion that use of commercial search engines was and is highly problematic and potentially unreliable, an alternative source of measurement was sought. However, the question of how do you measure something that is as large and ever expanding as the web was encountered. As the web got larger, the problem became more acute. Therefore, it was asked: could a target be selected that would allow for a more focused study, with manageable limits, in terms of both time and resources available? Furthermore, while in 2003, it seemed reasonable to think of the web as a series of standalone sites that contained information that could, given sufficient resources and time, be fully traversed and analysed, as least theoretically. By 2007, however, it became obvious that the Web 2.0 phenomenon was changing the nature of how information on the web was being stored and how users interacted with the web. Thus, a count of the number of 'pages' or 'sites' that were in a particular language would be one measure of the web, but an equally important measure would be to measure the effect of web 2.0.

Two possibilities emerged:

1) find an area or subset of webpages that could be said to be representative of the web; or

2) conduct a number of focused studies that together could be said to form a representative sample of the web.

4. Choosing Wikipedia as the Target

At this point it became obvious that Wikipedia's ability to be ported into a number of languages was a very encouraging development for minority languages, and therefore that Wikipedia was an interesting topic for minority language material on the web. The key aspects of Wikipedia from the point of view of this study are:

- Wikipedia is essentially organised by language, and not, as other parts of the web, by national domain or type of organisation. Each language edition of Wikipedia is essentially a complete website with its own home page, index and content and therefore can be treated as a single unit of study;
- At the time of writing, there were approximately 276 different language editions of Wikipedia;
- It is discrete, in that each language edition of Wikipedia has a home page, from which a complete index of the site is available;
- The individual language editions of Wikipedia are generally identical in terms of structure, layout and format; only the actual content differs;
- Wikipedia is a paradigm of Web 2.0; the content is both created and consumed by ordinary users.
- Unlike other parts of the web, including other Web 2.0 sites, most of Wikipedia is not hidden from normal users;
- Wikipedia itself produces a number of statistics;
- It does not require enormous resources to access. The English language edition of Wikipedia, at the time of writing, contained about 3 million articles. While it is not a trivial exercise to search such a dataset, it is well within the means of a single computer with only moderate storage;
- Wikipedia is open sourced and readily accessible and there are no major technical or legal reasons that prevent the full study of Wikipedia

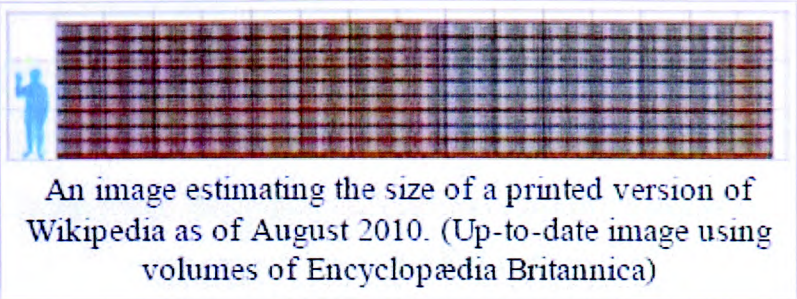


Figure 4 - The Size of Wikipedia

It was thought that Wikipedia would be a suitable, and indeed an ideal, candidate for measurement. Wikipedia is a discrete system that can be measured with some degree of certainty and accuracy. Following a presumption that, all other things being equal, two language communities with the same number of speakers would, over a period of time, produce roughly the same level of content in a given system, then an accurate measurement of the content generated by these language communities would allow us to test for presence.

Of course, no two languages are equal. Some languages have more speakers than others, and, with more members, more content would be written. Additionally, for cultural or educational reasons, some languages might tend toward greater content generation. However, it cannot be known *a priori* what the content of a particular language edition of Wikipedia is without first measuring it. Once some calculations have been made, and some reliable numbers and statistics are at hand, the hypothesis can be tested. Whether Wikipedia proves to be accurate measurement indicator, in its own right, or in conjunction with other indicators may be an important step in addressing the fuller question of how minority language presence on the web can be measured.

However, while Wikipedia is an excellent candidate for comparative language analysis, it is not a proposition of the present study that a measurement of Wikipedia is akin to measurement of the full web, or that Wikipedia can stand as a representative sample of the web. Wikipedia is a very unique concept and, for a variety of reasons, users of Wikipedia may not be normal users of the web. To foreshadow some of the results obtained from this study, it seems clear that some language communities are more 'fertile' than others, and this may be due to factors such as educational abilities, literacy and community activism.

However, it is hoped that by analysing a small part of the web, representative or unrepresentative, a theoretical model can be tested for basic soundness.

Furthermore, two possibilities of studying Wikipedia in isolation can be posited:

1) The study of Wikipedia as part of a general study of the web for minority language presence, and, used in conjunction with other studies, can be used to build up a bigger picture of minority language use on the web; or
2) It may be found, after a fuller study of the web, that Wikipedia is actually a representative sample of language provision on the web.

Until such fuller analysis of minority language is performed, there is no way of knowing how representative Wikipedia is. In the meantime, given the contribution that Wikipedia is making to minority language content on-line, the present study provides an interesting picture of the state of minority language editions of Wikipedia, when compared to majority language editions.

The ultimate goal, however, is not to measure Wikipedia *per se*, but to test a methodology for measuring web presence of a particular, especially minority, language. It is also suggested that Wikipedia could be used as a proxy indicator to measure the presence of any particular language on the web. Wikipedia is a unique concept. Unlike many parts of the web, but in common with other Web 2.0 applications, Wikipedia is user generated. Wikipedia may therefore provide a very good indicator of the interest of a particular community towards creating and using material written in their language. If a given language edition of Wikipedia has good coverage, a tentative conclusion could be drawn that that language may be have a good presence.

F. Outline of the Originality of this Study

While the target of this study is Wikipedia, the goal of this study is intended to be larger. While studying Wikipedia in isolation is an interesting question, there is no difficulty in understanding that the minority language editions of Wikipedia are quite small in both range of articles, as amply demonstrated by a study of Wikipedia's statistics in this regards, and, as any cursory consultation of most minority language editions will reveal, the depth of coverage is not great. It does not take a mathematical study to tell us what is easily ascertainable. The

goal of this study is to put forward a potential model for estimating and quantifying minority language material on the web. This principal originality of this study is not, therefore, in producing estimates of Wikipedia coverage, but rather a model, that uses Wikipedia as its test bed.

The principal originality of this study lies in the following:

- A definition of "presence" that is a two-dimensional concept combining both breadth of coverage and depth;
- A formula to measure the presence of a particular language on-line;
- A formula to compare the presence of one language with that of another;
- A "Language Constellation" system that measures languages in meaningful groupings based on real world competition model;
- A "tiered classification" model, that uses presence values to be predictive and descriptive of where a language's presence relates to other languages.

The principal goal of this study is to provide a framework for future work in the area. If it is clear what we is being measured, how it is being measured and how those measurements relate to other measurements, then a scientific study of language presence can be made.

As will be shown in the Literature Review, up until this point, all previous studies have attempted to measure primarily breadth only, and have not focused on depth of coverage. Most have used a model for measurement that calculates a relative proportion for each language studied, that, while interesting, does not give a clear picture of presence, and furthermore cannot be used to measure absolute growth. The model provided in this study provides a more complete picture of material available in each language, and provides an absolute measure as well as a comparative measure. Both of these measures are useful in that we can target a particular language, chart its growth over time, and, if necessary, make comparisons to other languages. This has not been possible with previous methodologies.

II. Literature Review

A. Studying Minority Languages on the Web

This study focuses on the question of how the presence of *minority* languages on the web can be measured. This subject has not been treated extensively in the existing literature, although there have been a number of studies that have measured *majority* language presence, and a small number of studies that have looked at individual minority languages and certain aspects of their use on-line. This body of work demonstrates a number of potential methods, and also highlights some of the pitfalls encountered when attempting web language measurement of any type, whether for majority or minority languages. Therefore, the following literature review will concentrate primarily on those studies that have attempted to measure majority language presence on the web. However, some care must be taken, since measuring majority languages poses a different set of questions, and a different set of problems, from measuring minority languages. The main problem is one of size and scale. Looking for objects that are in abundance is quite a different task than searching for objects that are difficult to find.

B. Taxonomy of Prior Methodologies

To date, there have been two fairly comprehensive reviews (Gerrand, 2007; Pimienta, Prada, & Blanco, 2009) of the studies that have been used to-date to measure language presence on the web. Gerrand and Pimienta classify the methodologies used to date as:

1) 'user profile' studies that measure the potential use of language by numbers of speakers of various languages;

2) 'user activity' studies that measure actual use of the internet and the language accessed by users;

3) 'web presence' studies that measure use of language on webpages or websites, with two subcategories:

a) studies using either the random IP sampling; and

b) studies using search engine results;

4) 'diversity index' studies that are a variation of web presence studies, but which use the diversity index proposed by (Greenberg, 1956) and deriving estimates of statistical variation;

5) 'undocumented studies' that purport to give figures, but no methodology is proposed);

6) miscellaneous studies using another methodology.

Not identified by these two authorsTo these we can add a number of other studies not classified by the either Gerrand or Pimienta:

7) studies that measure the webspace of a specific country, usually by analysing that country's Top Level Domain ("TLD) ('country specific TLD' studies); and

8) studies that only attempt to measure some aspect of one particular minority language ('targeted language' studies).

With the exception of numbers 1 and 2 above, user profile and activity studies, each of these methodologies will be discussed below. User profile studies and user activity studies are those that attempt to measure what and how users actually access on the web. This is a very interesting type of study, and may shed much light on what users actually do with the web, but a user profile study is the opposite of a 'web presence' study. Web presence measures what is available on the web and what is *potentially* accessible by a user; user activity measures what a user *actually* accesses (or perhaps what a user may want to access). Furthermore, there is a dramatic difference in how user profiles are measured from the other types of methodologies. User profile studies would require a test pool of actual users and a tracking system for analysing their web use, or some form of log, either at the client or server end. User studies thus measure human activity, the other studies measure the web as an artefact.

This literature review will only address the types of studies numbered 3 to 8 above, though in a slightly more compacted classification system.

C. Review of Prior Methodologies

1. Undocumented Studies

a) Global Reach (1996 - 2006)

A number of studies were conducted by an internet marketing company, Global Reach, from 1996 to at least 2005, which purported to show the number of internet users by language.[7] These studies have continually been referenced in the literature to date, even though the methodology and the results were somewhat suspect.

The following graph is a typical summary of the results of a typical Global Reach study:

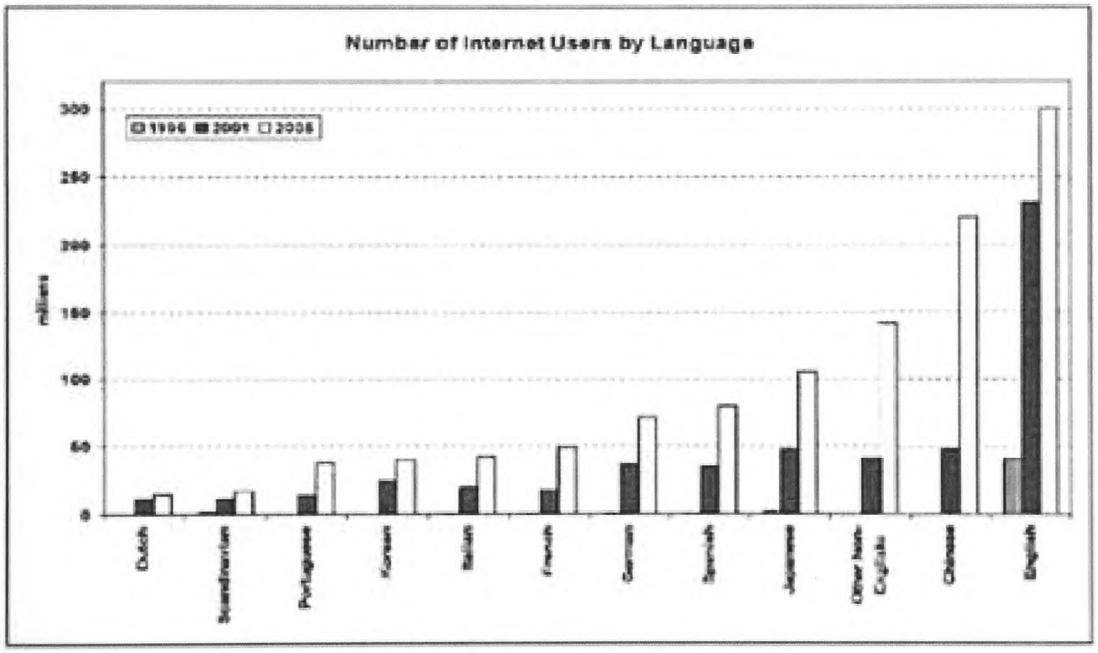


Figure 5 - Global Reach Results (2005)

These reports purported to show a breakdown of web presence by language (in this case Dutch, Scandinavian (sic), Portuguese, Korean, Italian, French, German, Spanish, Japanese, Other Non-English, Chinese and English). The methodology used was to extrapolate the

[7] http://www.global-reach.biz/globstats/index.php3, although this site is no longer operational

number of speakers per language that have access to the web and then calculate their presumed share of internet used. Gerrand deduced that they used an over-simplified formula to produce their statistics (Gerrand, 2007), as $i = \Sigma z(s(xyz) \,.\, (ayz \,/\, pyz))$. Where i is the number of internet users using a particular language; s is the number of first-language speakers of language x, in country z, in year y; ayz is the number of internet access services (dial-up, cable, etc.) available in a country in the year under study; p is the population of that country in the year of study.

The assumption behind this formula and calculation is that it was assumed that all persons who access the web do so in their first language. Needless to say, this assumption is unsound, especially when dealing with minority language users. As Gerrand points out, to assume that 12% of United States' internet users access the web in Spanish because 12% of the U.S. population speaks Spanish as a first language is simply not a valid assumption. Without some other rationale, it is impossible to know in what language users normally access the web. This would be even more difficult if the same logic were to be applied to minority language speakers, who often have good language proficiency in their relevant majority language.

Another, even more significant problem with Global Reach's methodology was that, while they purported to be an accurate picture of how users are using the internet, their studies do not actually monitor actual user use. They simply took estimated population statistics and extrapolated actual use of the web from those statistics. This is confusion between *potential* and *actual* use. More would have to be known about how users actually interact with the web and in what language before raw population figures can be extrapolated and draw conclusions therefrom.

It should also be noted that the Global Reach studies were not academic studies in the proper sense, but were produced for commercial reasons. This is likely to be the reason that the details of the methodology were not stated.

None of the languages that they produced calculations for included any minority languages, and as such the special difficulties inherent in this type of study were not addressed.

For this reason, the Global Reach studies, while often quoted in the literature are of little use in helping us determine a valid basis for language measurement on the web. In fact, it is

fair to say that the figures are nothing more crude estimates based on pseudo-science. At best, these estimates merely show that the web was initially taken up in the United States and other developed nations, and therefore that the majority languages of those states were the principal language of the internet in its first decade of existence.

2. Measuring by Random Sampling

The Random sampling methodology attempts to sample the entirety of the web by randomly creating a list of IP address, retrieving the data found at those addresses, and analyse the data returned. These types of studies were more popular when the web was in its infancy, but as the content available on the web increased dramatically these types of studies have declined. The majority of these studies have either attempted to measure the number of webpages, websites, number of bytes or other numerables (Babel Project, 1997; Crovella & Krishnamurthy, 2006; Pimienta & Lamey, 2001).

a) Babel Project (1997)

The earliest obtainable study of web presence by language was carried out by the Babel team, a joint initiative from Alis Technologies and the Internet Society, with the results published in 1997(Babel Project, 1997). The intent of the study was to give some indication of what languages were in use on the web at that time. This is a very early study and it should be noted that the web at that time was overwhelmingly a North American phenomenon.

The methodology used a random number generator to create an IP address. The IP address was fed into an ICMP protocol (ping) program that determined if a machine existed at that address. From a sample of more than 30 million potential addresses 60,000 machines were discovered. In a second step an additional program took each returned machine from the list, and determined if that machine hosted an HTTP server. More than 8,000 machines responded positively. A third step involved a linguistic analysis of the material available on the machines identified in the second step, using a language detection program – SILC - an online language identifier developed by the Laboratoire de recherche appliquée en linguistique informatique at the Université de Montréal.[8] The authors of this program claimed

[8] http://www-rali.iro.umontreal.ca/SILC/SILC.en.cgi

that it could identify a document's language and character set, and could recognize seventeen of the world's most used languages[9] in a wide variety of character sets. A fourth and final step consisted of the manual sampling of two hundred of the pages, and comparing the automatic detection with the visual identification. The authors stated that this manual sampling process was able to confirm the general reliability of the detection software and the process, but did reveal a number of flaws in the language detection software. The results of the study are detailed in following table:

Ranking	Language	Num Pages	Percentage	Corrected percentage
1	English	2 722	84,0 %	82,3 %
2	German	147	4,5 %	4,0 %
3	Japanese	101	3,1 %	1,6 %
4	French	59	1,8 %	1,5 %
5	Spanish	38	1,2 %	1,1 %
6	Swedish	35	1,1 %	0,6 %
7	Italian	31	1,0 %	0,8 %
8	Portuguese	21	0,7 %	0,7 %
9	Dutch	20	0,6 %	0,4 %
10	Norwegian	19	0,6 %	0,3 %
11	Finnish	14	0,4 %	0,3 %
12	Czech	11	0,3 %	0,3 %
13	Danish	9	0,3 %	0,3 %
14	Russian	8	0,3 %	0,1 %
15	Malay	4	0,1 %	0,1 %
	Unknown			5.6
Total		3 239	100 %	100 %

Table 2 - Babel Team Results (1997)

These results are not particular surprising given the embryonic state of the web in 1997. To the extent that they are valid results, they show that, in its infancy, the web was an overwhelmingly English language phenomenon.

The Babel team acknowledged a number of possible errors in their results. One problem identified was the possibility that a large number of machines could have been hidden behind firewalls that did not respond to a ping. Another source was caused by lost packets, a structural problem inherent in the TCP/IP protocol. Furthermore, they were only able to analyse the homepage of each server and did not allow them to uncover any further language

[9] English, German, Spanish, French, Portuguese, Italian, Chinese, Japanese, Serbo-Croatian, Danish, Dutch, Czech, Swedish, Norwegian, Finnish and Malay

options that could be accessed from the initial homepage. This would have had serious effects on the figures since it was, and may still be, a practice to present a homepage in one language with the possibility of one or many equal editions available in other languages. Given the early rollout of the web in North America, this may have had the effect of overestimating the amount of English, given that English is the primary language of that region and the tendency to write homepages in the majority language. They also admitted to problems in the language detection software, although these were not stated precisely.

Given that the authors noted a number of flaws in their software and methodology, this is nonetheless a useful study of the web in its very early days. It is unfortunate that the SILC program was not able to identify more languages, and more unfortunate still that no minority languages were identified. English had an initial head-start over other languages, and it would be interesting to see this study repeated at a later date to see whether, as the web develops and becomes stable, whether a marked decrease in the predominance of English on the web can be detected. It would be surprising if it were not.

As a methodology it is inherently sound if one were primarily concerned with the language of homepages. However, as stated above, multilingual content is often provided on otherwise unilingual webpages. Another drawback of this methodology is its inability to deal with Web 2.0. The homepage of a Web 2.0 sites tells us little to nothing about the content beyond the homepage. Given the increasing importance of Web 2.0 and dynamically generated web content, this methodology is unlikely to yield accurate and reliable results of any language use on the web, and in particular is unlikely to be of much us as the web continues to develop more dynamically created content.

b) Lavoie & O'Neill (1999)

Another study, conducted in 1999, by Office of Research Web Characterization Project ("OCLC") (Lavoie & O'Neill, 1999) also used the random sampling technique. The scope of this study was to look at two particular aspects of the web: 1) the country of origin of each website's publisher and 2) the language used by the websites.

Their methodology was to use a specially configured random number generator, and using a 0.1% random sampling (without replacement) of the Internet Protocol address. This yielded a list of 4,294,967 unique, random IP addresses. For each of these IP addresses, an HTTP connection was attempted. If the IP address returned an HTTP response code and a

document in response to the connection attempt, then the address was identified as a website. Each website identified in the sample was harvested using specially developed software. Once the website data was collected several diagnostic tests were applied to identify sites duplicated at multiple IP addresses. This yielded an estimate of the total number of unique websites. For the language analysis, manual inspection was generally confined to the homepage. If the content of the homepage was in a single language, and no references were made to content elsewhere on the site available in other languages, the site was considered monolingual. To identify non-English languages, an improved version of the software used by the Babel Project (SILC) was used. In this version twenty-nine languages were claimed to be identifiable by the study.[10] The results of the study were:

Language	% of 1999 Sites	% of 1998 Sites
English	79.7	83.7
German	7.5	7.9
Japanese	3.6	3.2
French	3.2	3.0
Spanish	2.8	2.2
Portuguese	2.6	2.1
Italian	2.0	1.3
Chinese	1.7	1.0
Dutch	1.2	1.4
ALL OTHERS	1.0	1.0

Table 3 - Lavoie & O'Neill Results (1999)

As in the previous study, it can be seen that web was mostly an English language affair in the first five or six years. However, it can be noted that if these two studies paint an accurate picture, the web moved from being 83% in English to 79% within only two years, which would indicate both that English's position as the overwhelming language of the web in its early years was a temporary phenomenon.

[10] Arabic, Greek, Polish, Chinese, Hebrew, Portuguese, Czech, Hungarian, Russian, Danish, Icelandic, Serbo-Croatian, Dutch, Italian, Slovenian, English, Japanese, Spanish, Estonian, Korean, Swedish, Finnish, Lithuanian, Thai, German and Norwegian

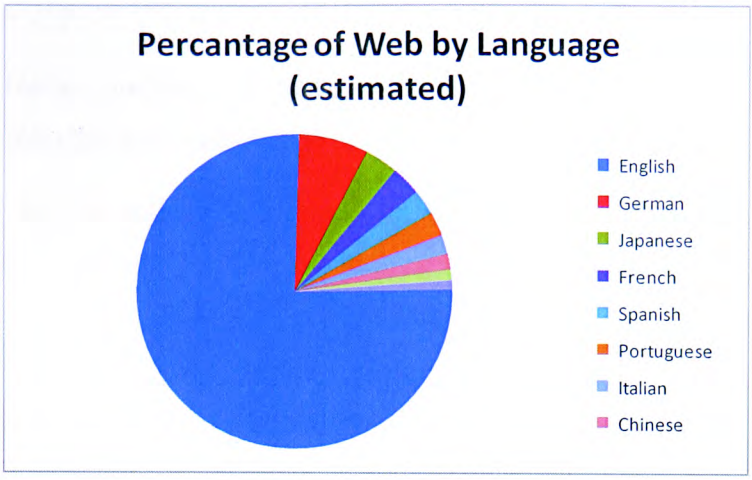


Figure 6 - Lavoie and O'Neill (1999)

The authors were able to conclude that the web at that time (1999) was exhibiting signs of an "ongoing internationalization process", that the distribution of the countries of origin of websites was widening, and the share of websites provided by sources in the English-speaking world was decreasing. The number of languages identified in the 1999 increased substantially as compared to the 1997 Babel project study (Lavoie & O'Neill, 1999).

c) O'Neill et al. (2003)

Lavoie and O'Neill (along with Bennett) repeated their attempt to analyse the public web again in 2003 (O'Neill, Lavoie, & Bennett, 2003). They adopted a similar methodology as previously and reached the following results:

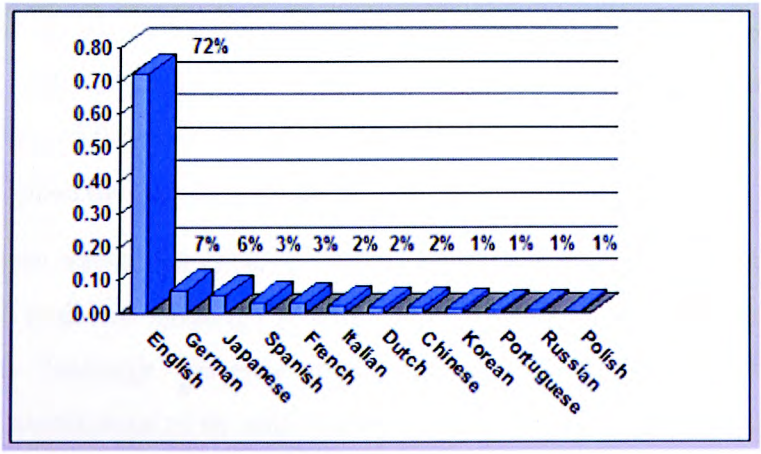


Figure 7 - Lavoie & O'Neill (2002)

They were able to conclude that "[t]he ubiquity of the public Web in other parts of the world has not reached the level realized in the United States" (O'Neill et al., 2003).

As is noted by the authors the study's main weaknesses lay in the abilities of the language detection software, which was limited to twenty-nine languages. However, no details are given as to the language recognition algorithms. Secondly, there was only one sample taken for each year. It can also be noted that the results for only twelve languages were reported and these accounted for 101% (due, no doubt to rounding off errors) of the web. We can note that some languages that have large numbers of speakers (notably Arabic and Indonesian) are missing from the study, and some important European state languages are likewise unaccounted for (Danish, Norwegian, Swedish, Czech, Greek, Finnish, etc.). Lastly, no minority languages were included. Granted that these might be a very small fraction of the web, but some traces of these languages would be expected to be returned in any meaningful sampling of the web. Had this study been more robust, it might have given us a useful snapshot of the web at the time of the study, but given these limitations, the study mainly serves as a demonstration that English is very prevalent on the web from 1999 to 2002.

Like the previous Babel study, the two Lavoie and O'Neill studies used inherently sound methodologies and provided interesting snapshots of the linguistic situation of the web in its infancy.

d) Critique of Random Sampling Methodology

The three studies detailed above provided some very early snapshots of the web in its infancy. As was to be expected, the results showed that English was in an indisputably dominant position, with only minor use of the other major languages. This was expected at the time, since uptake of the web was initially focused in North America. Undoubtedly, the picture will change when the web is fully globalised.

Besides the above mentioned critiques of the random sampling methodology as it relates to measurement of majority language presence, any further use of this methodology for measuring minority language presence must address four problems. Firstly, language detection software would have to be enlarged to include the minority languages that are the target of any potential study. Secondly, since minority language material is by definition rarer

than majority language material, the sample size would have to be increased even further, to ensure that minority language material is not missed simply because the sample size is too small. Thirdly, the size of the web has grown exponentially since these studies, with a likely result that more sites would have to be interrogated to produce meaningful results. Fourthly, structural changes in the web would have to be taken into account, notably to take into account the Web 2.0 problem).

If these problems were sufficiently addressed, such a study could provide useful results for the measurement of minority language presence. Indeed, by using very large sample sizes and very accurate language detection algorithms, such a study could produce excellent results. But the key lies in these two problems. Large samples take a long time to collect and a long time to verify. Writing and debugging the language detection algorithms for a large number of languages is a non-trivial exercise that requires both linguistic and programming skills. It can easily be foreseen that such an exercise, if performed in 2010, would require many millions of hits and some very sophisticated language detection software. As the web gets even larger, these requirements become even more extensive.

3. Web Presence Measured by Search Engine Results

Since random sampling of a very large object such as the web requires very large samples, with the attendant costs in terms of money, time and storage and computational requirements, it is no surprise that others have sought more cost and resource efficient methods. One such method is to use already existing search engines. Since search engines have already scraped a large part of the web and made the results of such scrapes available along with some additional information, the hard part of the data collection task is completed, or so runs the argument. As was discussed above (in section I.E of the Introduction), search engines were used extensively in the early stages of this study as being an ideal manner of obtaining the data for the calculation of minority language presence.

a) Grefenstette & Nioche (2000)

Grefenstette and Nioche conducted a large and comprehensive survey of the web between 1996 and 2000 (they conducted three separate studies in October 1996, August 1999, and February 2000) (Grefenstette & Nioche, 2000). They stated used a rather novel approach than what later studies would use: rather than estimating the number of webpages, they attempted to estimate the number of words that were contained in a particular language within a search

engine's database, and then by comparing the total number of words for each language, they estimated the relative proportion the languages that were chosen for their study.

Their methodology consisted of obtaining a 1 megabyte training text, from which they extracted all alphabetic tokens. Then the frequency of these alphabetic tokens were ascertained within the training text, and the list of tokens was sorted by descending frequency to determine the 100 most frequently occurring tokens. In order to avoid homographic[11] confusion they eliminated from each list any word appearing in more than one list. From the remaining tokens, they retained the top twenty most frequently occurring tokens and their frequencies were then used as predictors for that language (Grefenstette & Nioche, 2000:3).

In the next stage of the process the files were tokenised for each language, using non-alphabetic characters as separators, and the tokens were sorted and counted for frequencies in a new file, for each predictor. The authors then divided the frequency of that token by the relative frequency of the predictor, producing that predictor's estimate of the total number of words. After throwing out the two highest and the two lowest estimates, they averaged the remaining predictions and gave the average as the prediction of number of words in the new file for the language being estimated (Grefenstette & Nioche, 2000:4).

As a final stage, they used the web search engine AltaVista (now defunct) and formed a query composed of the predictor words and obtained two counts for each query using the twenty predictors for each language. AltaVista responded with a numerical result for each predicator word. After dividing this actual frequency by the relative frequency, this lead to a single-word prediction. For example, the German word '*oder*' had a determined relative frequency of 0.0056118 in their training set and '*oder*' was found 13566463 times within the Alta Vista database. By this method, they calculated that a total of 2,417,488,684 German words were accessible through Altavista. By doing the same analysis for all 20 predictors, and throwing out the highest and lowest results, they were able to estimate the total number of words searchable in Altavista. The ratio of seven languages to English is given for each of these three time periods, as shown in the following table:

---

[11] This is when a word exists with the same spelling in two or more different languages. For example, a word '*que*' exists in a number of Romance languages. It may often be the same word in a language, or it may be different (e.g. '*gift*' in English and German). The existence of these homographs has the ability to greatly affect language recognition algorithms.

Language	Words Oct 1996	Ratio to English	Words Aug 1999	Ratio to English	Words Feb 2000	Ratio to English
English	6,082,090,000	1.000	28,222,100,00	1.000	48,064,100,00	1.000
German	228,938,428	0.038	1,994,229,409	0.071	3,333,127,671	0.069
French	223,316,023	0.037	1,529,795,169	0.054	2,732,221,327	0.057
Spanish	104,319,158	0.017	1,125,646,460	0.040	1,894,966,981	0.039
Italian	123,555,682	0.020	817,270,444	0.029	1,338,351,674	0.028
Portuguese	106,167,245	0.017	589,391,943	0.021	1,161,898,076	0.024
Norwegian	106,497,066	0.017	669,331,120	0.024	947,486,593	0.020
Finnish	20,647,404	0.003	107,260,274	0.004	166,599,467	0.003

Table 4 - Greffenstette Results (1996 - 2000)

Table 5 below gives the raw numbers of the study as reported in August 2000, with additional columns that show the calculated ratios and percentages for each language. What is especially notable from these two tables is the high proportion English with the Alta Vista database, and, by the authors' logic, on the web. As stated before, and noted by other studies at approximately that time, this was an early period in the life of the web and access was primarily located in North America and the West.

Language	Wordcount Estimate	Ratio to English	%
English	47,264,700,000	1.00000	0.78518
German	3,068,760,000	0.06493	0.05098
French	2,208,418,000	0.04672	0.03669
Spanish	1,595,489,000	0.03376	0.02650
Italian	1,240,205,000	0.02624	0.02060
Portuguese	924,965,000	0.01957	0.01537
Swedish	644,740,000	0.01364	0.01071
Dutch	622,063,000	0.01316	0.01033
Norwegian	453,391,000	0.00959	0.00753
Czech	269,310,000	0.00570	0.00447
Hungarian	268,944,000	0.00569	0.00447
Polish	235,726,000	0.00499	0.00392
Danish	206,167,000	0.00436	0.00342
Finnish	192,105,000	0.00406	0.00319
Slovakian	140,909,000	0.00298	0.00234
Catalan	126,324,000	0.00267	0.00210
Malay	113,236,000	0.00240	0.00188
Turkish	100,548,000	0.00213	0.00167
Slovenian	74,998,000	0.00159	0.00125
Croation	72,122,000	0.00153	0.00120
Roumanian	63,846,000	0.00135	0.00106
Icelandic	53,167,000	0.00112	0.00088
Irish	49,778,000	0.00105	0.00083
Estonian	43,257,000	0.00092	0.00072
Latin	38,256,000	0.00081	0.00064
Basque	28,296,000	0.00060	0.00047
Esperanto	26,795,000	0.00057	0.00045
Latvian	21,925,000	0.00046	0.00036
Lithuanian	20,927,000	0.00044	0.00035
Breton	9,975,000	0.00021	0.00017
Albanian	9,203,000	0.00019	0.00015
Welsh	7,590,000	0.00016	0.00013

Figure 8 - Greffenstette Results

If the results of the above table are graphed, we can clearly see the overwhelming presence of English as shows by these studies:

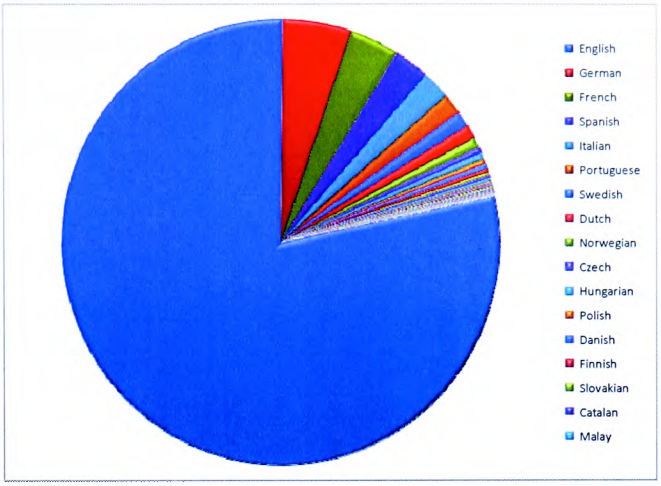


Figure 9 – Grefenstette Results

It is interesting that this is one of the few studies that included some minority languages (Irish, Breton and Welsh) and a regional language (Catalan). If the details for the three Celtic languages are extracted it can be readily seen from this study that Irish is much more prevalent on the web than the other two Celtic languages:

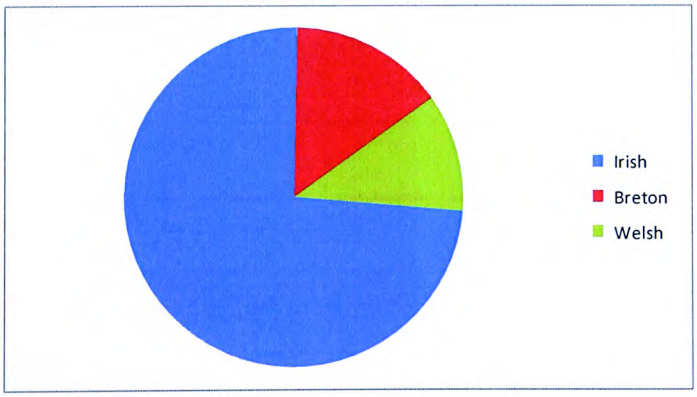


Figure 10 – Grefenstette Pie Results (Celtic Languages)

In general this was an interesting study that used an interesting methodology, and furthermore attempted to look at a range of languages beyond the usual widely spoken

languages. There are a number of critiques that can be made with respect to this study. The most important of which was the reliance on the search engine methodology. A more detailed critique of this will be discussed in more detail below. As mentioned by the authors, at the time AltaVista only indexed about 16 percent of the web, and it is impossible to know whether the language distribution across the entire web is reflective of this one portal's indexing coverage. Thus, it can be argued that Alta Vista is not properly representative of the web. It can also be asked whether there was any bias within the Alta Vista system to favour English language websites and web pages. Without full access to Alta Vista's algorithms and search strategies, it would be impossible to know how much bias there was. However, this was undoubtedly a serious study and it would have been interesting to see it repeated after the year 2000. Since Alta Vista no longer exists, it is impossible to repeat these results. Of course, the methodology could be easily applied to another search engine.

b) Large and Moukdad

In June 1999 Large and Moukdad analysed the returns of pages indexed by the search engine AltaVista (Large & Moukdad, 2000). Their method was a simple one: by using the AltaVista interface, they selected one of the languages from Altavista's menu, and then entered into the search box a meaningless string, preceded by the negation operator. This caused AltaVista to return all the results which did not contain the search string, together with a count of such results. From this, the authors were able to calculate the number of pages indexed in the AltaVista database, for all the languages which AltaVista allowed. Their results are as shown in the following table:

Ranking	Language	Indexed pages
1	English	198,623,158
2	German	20,101,601
3	Japanese	5,265,839
4	French	4,889,844
5	Spanish	4,083,809
6	Swedish	2,539,036
7	Italian	2,523,000
8	Portuguese	2,405,744
9	Dutch	2,346,680
10	Chinese	1,731,619
11	Danish	856,268
12	Czech	852,778
13	Polish	792,194
14	Russian	582,898
15	Korean	566,451

Table 5 - Large and Moukdad Results (1999)

These results were comparable with those from the Babel Team. The relative positions of the top nine languages are nearly identical; only Chinese, Polish, and Korean are absent from the Babel Team study. Although English is the most prevalent, nine other languages total more than 1.5 million pages each.

The authors noted a number of problems with their methodology: 1) search engines do not index exactly the same pages, and the results of an analysis of AltaVista would not be identical to an analysis of other search engines; 2) given the rate of growth of the web, the results were likely to change over time; 3) AltaVista did not permit a search of all the languages which it indexes (e.g. Arabic, for example, was indexed by Alta Vista, but was not searchable from the interface); and 4) AltaVista counted a page as being in a particular language even if it contained material in other languages. In other words, some pages may have been counted multiple times in different languages.

Large and Moukdad concluded that while English was dominant, many other languages, nevertheless, had a presence and that the proportion of English webpages was declining compared with those in other languages.'(Large & Moukdad, 2000:84)

What is most interesting about this study is the fact that it tended to mirror the Babel team results, but used a much simpler and much easy methodology. One could easily see that if Alta Vista were a valid benchmark for web presence, then this methodology would be preferable, especially given that its results were corroborated by a previous study.

Of course we could also argue that each of the studies was equally flawed. It may have been possible that English so prevalent owing to a tendency of US based search engines to favour US based websites, or that search engines might show some bias against certain languages, particularly those that used non-alphabetic systems.

c) Guinovart (2003)

Guinovart conducted a study in 2003 that used the AltaVista search engine to obtain the number of documents contained in that search engine's database. At the time AltaVista allowed for searching in a number of minority languages. The results of the study are presented in the following table:

Rank	Language	Pages	% total	Rank	Language	Pages	% total
1	**English**	**442M**	**60.727**	24	Slovenian	681K	0.093
2	German	51.2M	7.035	25	Greek	648K	0.089
3	Japanese	43.2M	5.926	26	Indonesian	597K	0.082
4	Chinese	26.2M	3.600	27	Ukrainian	588K	0.081
5	French	24.6M	3.379	28	Croatian	521K	0.071
6	Korean	20.4M	2.799	29	Hebrew	514K	0.071
7	Russian	19.6M	2.689	30	Icelandic	441K	0.061
8	**Spanish**	**16.4M**	**2.254**	31	Romanian	419K	0.058
9	Italian	15.1M	2.077	32	Arabic	328K	0.052
10	Portuguese	12.5M	1.718	33	Lithuanian	328K	0.045
11	Dutch	11.2M	1.533	34	Bulgarian	319K	0.044
12	Polish	7.46M	1.024	35	Malay	194K	0.027
13	Swiss	6.56M	0.900	36	Latvian	156K	0.021
14	Czech	5.94M	0.815	**37**	**Galician**	**99.0K**	**0.014**
15	Danish	5.03M	.690	**38**	**Basque**	**80.3K**	**0.011**
16	Norwegian	3.64M	0.499	39	Afrikaans	71.9K	0.010
17	Finnish	3.02M	0.414	40	Vietnamese	48.4K	0.007
18	Slovakian	1.97M	0.270	41	Byelorussian	43.6K	0.006
19	Hungarian	1.91M	0.262	42	Welsh	42.7K	0.006
20	Turkish	1.51M	0.208	43	Faroese	37.3K	0.005
21	Thai	855K	0.117	44	Albanian	33.0K	0.005
22	Estonian	811K	0.111	45	Friesian	21.0K	0.003
23	**Catalan**	**681K**	**0.094**		*All the above*	*729M*	*100.00*

Table 6 - Guinovart Study (2003)

The results look impressive and based on *a priori* expectation, would seem to paint a realistic picture of the web of 2003. Notably English comprises 60% of the web, a lower figure from what was earlier reported, and an expected result considering the increasing use of

the web in other parts of the world. As noted above, the relative dominance of English has declined in each study. German was, as previous studies showed, the second most prevalent language, with Japanese a close third. Again, previous studies lead us to expect this result. Other notable results are that European languages (i.e. those languages that have their origin in Europe, irrespective of where they are spoken) account for 35 of the 45 languages reported.

However, further study shows some significant missing elements. Hindi, Urdu and Farsi are not accounted for. Welsh is included, but not Irish or Breton. Croatian is listed but not Serbian. What the "Swiss" language is, is not properly explained[12]. At the time, AltaVista only allowed searches in forty languages.[13]

The principal defect in this study, as in the previous studies, is that the results are completely dependent upon AltaVista and therefore the conclusions of this study are subject to the general criticism of all search engine based results, that is more fully addressed below.

d) Mas (2003)

Mas conducted a similar study in 2003, but using the now defunct AllTheWeb search engine (Mas i Hernàndez, 2003). His methodology used term frequencies from which he obtained the raw results returned by that search engine. He admitted that such a methodology was not entirely trustworthy, though in his view this mainly arose with respect to pages with little content or when results were returned from closely related languages (e.g. Occitan and Catalan), that caused homographic interference. The results of his study were as follows (with percentage values added to show the relative percentage for each language):

[12] Presumably 'Swiss' is German, but that would be a remarkable fact that it ranks so highly

[13] Albanian, Arabic, Bulgarian, Catalan, Chinese (Simplified), Chinese (Traditional), Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hebrew, Hungarian, Icelandic, Indonesian, Italian, Japanese, Korean, Latvian, Lithuanian, Malay, Norwegian, Persian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish, Tagalog, Thai, Turkish and Vietnamese

Language	Number of Pages	%		Language	Number of Pages	%
English	1,280,041,397	60.42%		Greek	2,366,733	0.11%
German	182,005,546	8.59%		Romanian	2,052,990	0.10%
French	99,737,704	4.71%		Slovenian	1,685,426	0.08%
Japanese	69,730,375	3.29%		Croatian	1,672,582	0.08%
Spanish	65,814,567	3.11%		Estonian	1,464,539	0.07%
Chinese	65,730,212	3.10%		Irish	1,391,302	0.07%
Korean	64,606,324	3.05%		Bulgarian	1,120,713	0.05%
Russian	42,276,247	2.00%		Lithuanian	1,078,341	0.05%
Italian	41,849,365	1.98%		Indonesian	1,044,038	0.05%
Dutch	41,119,851	1.94%		Ukranian	1,014,537	0.05%
Portugese	37,695,762	1.78%		Estonian	559,913	0.03%
Polish	22,154,325	1.05%		Byelorussian	535,697	0.03%
Czech	15,580,583	0.74%		Vietnamese	390,164	0.02%
Swedish	14,901,968	0.70%		Malay	327,947	0.02%
Danish	12,107,133	0.57%		Galician	274,132	0.01%
Hungarian	8,540,941	0.40%		Basque	154,709	0.01%
Norwegian	8,123,301	0.38%		Latvian	137,355	0.01%
Finnish	5,678,599	0.27%		Afrikaans	115,689	0.01%
Slovakian	5,077,965	0.24%		Welsh	93,027	0.00%
Hebrew	4,792,646	0.23%		Faroese	65,785	0.00%
Turkish	4,704,212	0.22%		Frisian	63,236	0.00%
Thai	3,124,572	0.15%		Albanian	53,236	0.00%
Catalan	2,926,550	0.14%		Serbian	42,848	0.00%
Arabic	2,470,616	0.12%		Swahili	14,314	0.00%

Table 7 - Mas Results (2003)

Which can be more readily seen if we present them in a pie chart:

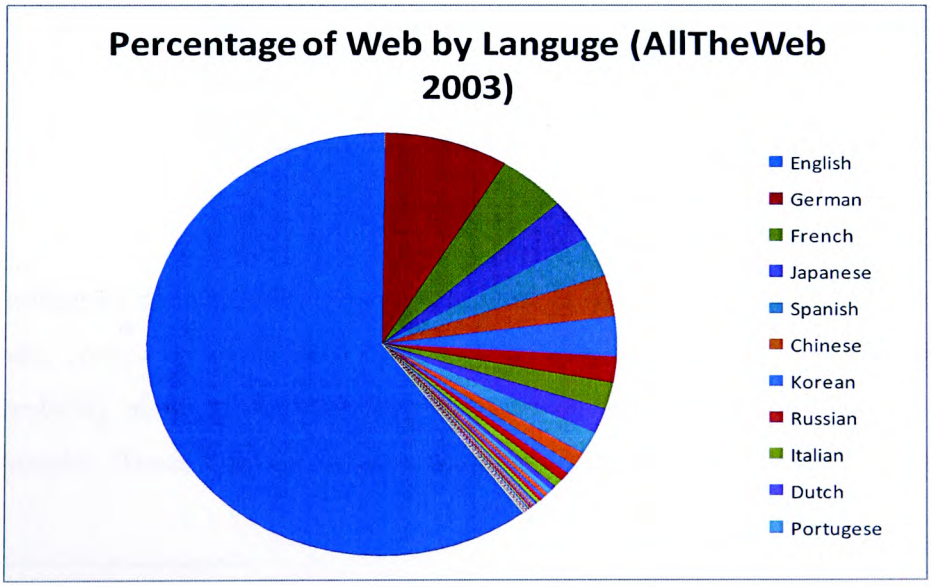


Figure 11- Mas Results (2003)

Mas then reordered some[14] of the results, using estimated populations (using the then current Ethnologue estimates) of the speakers of such languages on a "page per person" basis, to produce the following comparison:

Language	Pages	Population	Page/Person
Icelandic	1,391,302	250,000	5.57
English	1,280,041,397	341,000,000	3.75
Danish	12,107,133	5,326,000	2.27
Dutch	41,119,851	20,000,000	2.06
German	182,005,546	100,000,000	1.82
Swedish	14,901,968	9,000,000	1.66
Norwegian	8,123,301	5,000,000	1.62
French	99,737,704	72,000,000	1.39
Estonian	1,464,539	1,100,000	1.33
Czech	15,580,583	12,000,000	1.3
Finnish	5,678,599	6,000,000	0.95
Hebrew	4,792,646	5,150,000	0.93
Slovakian	5,077,965	5,606,000	0.91
Slovenian	1,685,426	2,000,000	0.84
Korea	64,606,324	78,000,000	0.83
Italian	41,849,365	62,000,000	0.67
Hungarian	8,540,941	14,500,000	0.59
Japanese	69,730,375	126,000,000	0.55
Polish	22,154,325	44,000,000	0.5
Catalan	2,926,550	6,565,000	0.45
Russian	42,276,247	167,000,000	0.25
Portugese	37,695,762	176,000,000	0.21
Spanish	65,814,567	322,000,000	0.2
Greek	2,366,733	12,000,000	0.2
Thai	3,124,572	20,047,000	0.16
Croatian	1,672,582	21,000,000	0.08
Romanian	2,052,990	26,000,000	0.08
Turkish	4,704,212	61,000,000	0.08
Chinese	65,730,212	885,000,000	0.07
Arabic	2,470,616	213,223,637	0.01

Table 8 - Mas - Languages by Population (2003)

As admitted by the author, the methodology chosen was somewhat suspect. We can note that Icelandic comes in first position, by quite a margin over English, in this population analysis, probably owing to the fact that Icelandic is only spoken by some three hundred thousand people. These type of exaggerations are quite common when dealing with small

[14] The author did not give reasons why he only made calculations for of the languages. It seems he was only interested in the major languages and some of the European regional languages. It seems that his main interest was Catalan.

populations and comparing them to the results for large populations. For example, if a language were spoken by only two persons, and that one of them wrote 12 webpages, that language would instantly become the most spoken language on the web 'by population'. As is shown in the above chart, Icelandic is the largest language on the web 'by population'. Is it because Icelandic speakers are almost twice as productive as English speakers? Or is it because the extremely small population of Icelandic speaker relative to English that exaggerates the results? And even if Icelandic were more productive on a per person basis, does that necessarily translate into a meaningful basis of comparison? Again, small, but relatively production language communities may do well on such a comparison, but if the total output does not cover a broad spectrum, we can question the point of producing this kind of analysis.

His study is nonetheless interesting in that it attempts to reorder the rankings of languages by numbers of speakers. This is not necessarily wrong, but it is susceptible to some surprising results. This study will specifically address 'per population' analyses below.

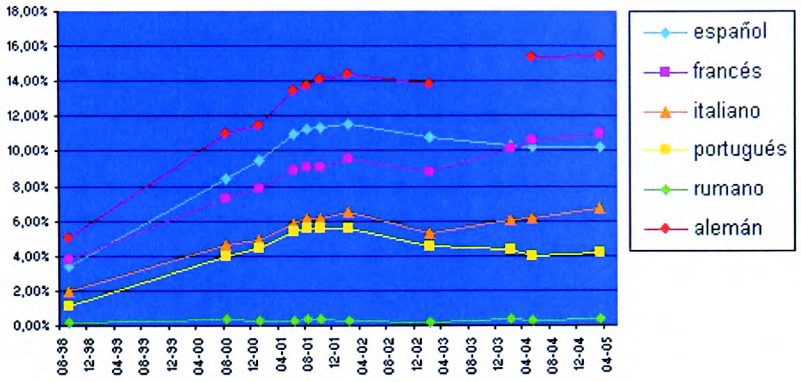
e) The FUNREDES Studies (1998 – Present)

The FUNREDES organisation has conducted a number of studies from 1996 to about 2008 (see http://fundredes.org). It has almost exclusively concerned itself with analysing and comparing the situation of the five most spoken Romance languages (Spanish, French, Portuguese, Italian and Romanian) and has attempted to compare these with English and German. As such, the studies are not directly applicable to minority languages, but the methodology is interesting.

The FUNREDES method, though continually refined, consists of using search engines to obtain the number of occurrences of a given word in a given sector of the web. A sample of keywords in each language is constructed, with particular care taken to providing the best semantic and syntactic equivalence among the different languages. These results are then analysed using traditional statistical tools. The test is rerun at different intervals (Pimienta, 2005: 32).

Generally, the various studies showed a fairly positive increase in the use of the target Romance languages on the web from 1998 to 2005. Figures released by FUNREDES showed that in 1998 the percentage of pages in any of the six target languages (the five studied Romance languages and German) was 15.41% of the web's total (with the remainder being in

English and all the other languages), whereas this proportion had increased to 47.94% in 2005.



	09-98	08-00	01-01	06-01	08-01	10-01	02-02	02-03	02-04	05-04	03-05
español	3,37%	8,41%	9,46%	10,95%	11,24%	11,36%	11,60%	10,83%	10,30%	10,19%	10,23%
francés	3,75%	7,33%	7,89%	8,86%	9,13%	9,14%	9,60%	8,82%	10,18%	10,64%	11,00%
italiano	2,00%	4,60%	4,93%	5,88%	6,15%	6,15%	6,51%	5,28%	6,09%	6,15%	6,77%
portugués	1,09%	3,95%	4,44%	5,40%	5,57%	5,61%	5,62%	4,55%	4,36%	4,02%	4,15%
rumano	0,20%	0,37%	0,33%	0,32%	0,35%	0,36%	0,33%	0,23%	0,41%	0,31%	0,37%
alemán	5%	11,00%	11,43%	13,42%	13,74%	14,08%	14,41%	13,87%		15,37%	15,42%

Figure 12 - FUNREDES Figures for 1998 to 2005

The relative use of English had declined dramatically in the six and a half years of the study. Assuming for the moment that English accounted for the entirety of the residual web not accounted for in the above figures, it can be seen that English declined from something in region of 84% of the web to under 53% (FUNREDES, 2005).

While there may have been some alarm, especially in 1998, as to the low relative percentage of some of the figures, it is easy to see the chief reason in the low proportions of the web written in the target Romance languages being the initial rollout of the web in North America. As the figures show, there is was a healthy increase in the use of the other major languages. Indeed, French and Italian in particular show, in 2005, a relative position that far exceeds their relative proportion of the world's language users.

This study has been an on-going for many years. The most recent results were issued in 2009 (Pimienta et al., 2009)

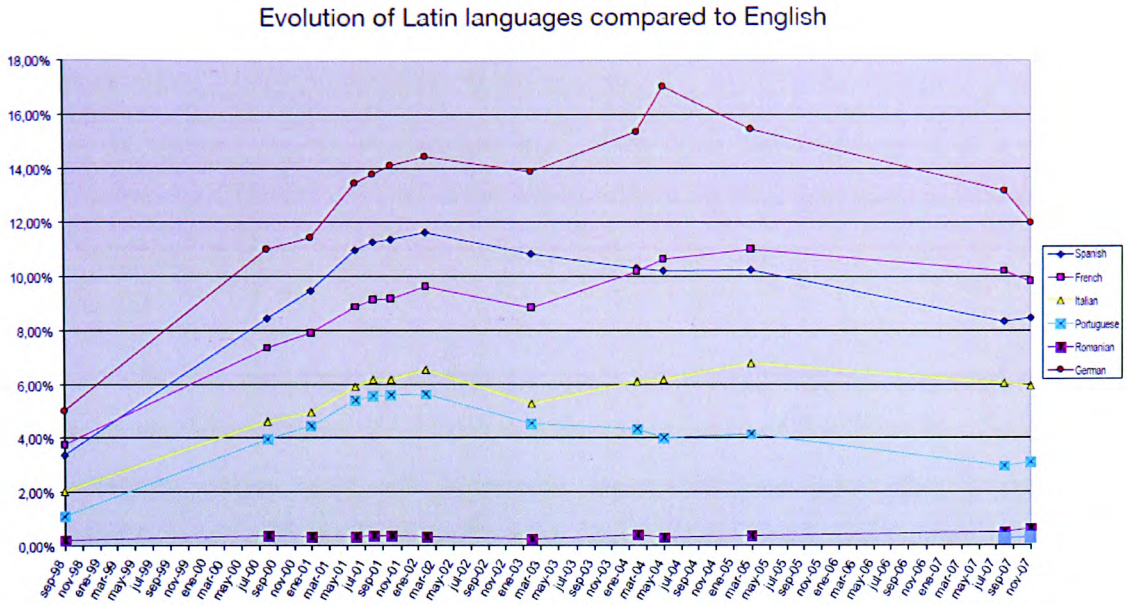


Figure 13 - FUNREDES Results 2007

Pimienta admits several weaknesses in the method, including: 1) the inability to provide only relative figures (i.e. Spanish compared to English) and not absolute figures (i.e. the size of the Spanish web); 2) the study cannot easily incorporate new languages; 3) it only looks at the public web and does not take into account the 'invisible' web; and 4) it is dependent on the accuracy of search engines (Pimienta, 2005: 32). Indeed, the authors noted that there were considerable problems with the use of search engines:

> "The whole process of this study has been characterized by a permanent struggle with search engine behaviours. The primary activity for each measurement campaign was to validate whether the search engines could meet the methodological requirements of the research, and, in many cases, understand the rationale for what first appeared as invalid results." (Pimienta et al., 2009: 20)

In the end, the authors were confident that they had solved the major issues.

As will be discussed below, the use of search engines for producing the raw data for the study of the web is quite problematic.

f) "Authenticity Test" Using Google Search Results

Kelly-Holmes (Kelly-Holmes, 2006) developed a method for measuring Irish language use on the internet: (1) an "authenticity test" of the Irish language version of a search engine like Google. 'Authenticity' in this case means obtaining Irish language results is response to Irish language queries; and (2) a measurement of the range of domains in which Irish is represented.

For this test, the version of Google found at www.google.ie was used and the Irish language interface selected. Five Irish words, one for each of the domains of arts, education, entertainment, native sport and commerce: *banc* ('bank'), *siopa* ('shop'), *ceol* ('music'), *airgead* ('money') and *peil* ('football') were entered into Google and the first ten hits returned were analysed.

The analysis was on a sliding scale from a fully monolingual Irish language page or site to a monolingual English language site. The categories were 'IR' (monolingual Irish), 'BI' (fully bilingual), 'Ie' (mainly Irish, with minor borrowings of English), 'EI' (mainly English with some Irish words), 'Ei' (mainly English with only symbolic use of Irish) and 'EN' (monolingual English).

Figure 14 below shows the results of the test for the Irish language for searches in August 2004. In total, 69% of the hits returned were 'valid' Irish language hits (IR, IE, Ie), which is impressive, although it should be clearly stated that the test involved entering Irish language terms into the Irish version of Google and selecting the "pages from Ireland" option.

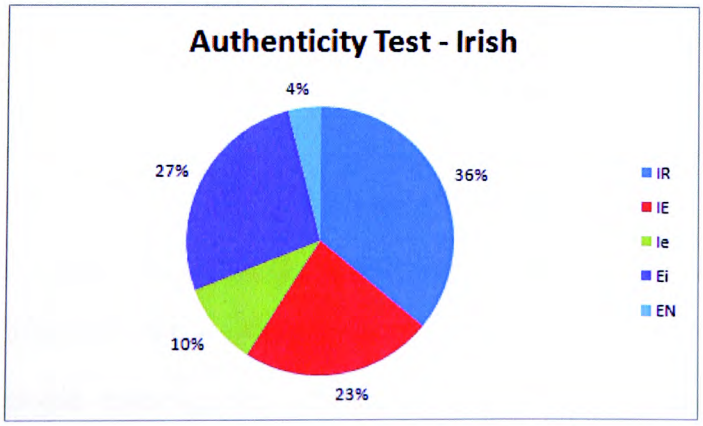


Figure 14 - Authenticity Test - Irish

It should be noted that the data sampling size was very small, only five samples. However, the principal difficulty with this test is that it relied on a localised version of Google (google.ie), which allowed for the searching of terms in either English or Irish. Since Google does not produce such localised versions for all other languages, it is difficult to replicate this authenticity test for other languages, even comparatively similar Celtic languages. For example, there is, at the time of writing, no Welsh language version of Google that allows for the searching of Welsh terms. Most of the world's minority languages are likewise ill served by Google. Furthermore, this is not particularly a 'web presence' study. It really does not tell us how much material is available in a particular language, and more relates to whether Google is returning Irish language results for Irish language queries. This study has only been included for completeness and because it was one of the few studies encountered that attempted any sort of measurement of a minority language.

g) Critique of Studies Using Data Derived From Search Engines

The web is growing exponentially and has long passed the point where a full traversal is anything but an enormous exercise that requires a large amount of money and time. As of January 2005, the web was estimated to have more than 11 billion pages (Gulli & Signorini, 2005). Given that the internet is vast and that minority language sites constitute, by definition, only a small portion of the web, commercial search engines can provide a more cost effective and more targeted form of data mining. Since existing search engines have already performed the hard task of scraping and indexing vast amounts of data at a huge cost, it would seem to be an obvious solution to use such search engines to do the hard part of the task. But, this has proved to have been problematic (Bolshakov & Galicia-Haro, 2003; Cafarella & Etzioni, 2005; Gerrand, 2007) .

A number of studies have been based on data obtained from search engines. Gerrand, in particular, has pointed out the deficiencies in using search engines in his study of linguistic diversity (Gerrand, 2007). In particular he notes the following problems:

1. Search engines have performed only limited sampling of the web. This relates to flaws inherent in the crawling of data;
2. Undisclosed indexing algorithms, making it impossible to know what was searched and how representative the contents of the;

3. Language search restrictions - few search engines allow for searches to performed in any but the major languages.

Since none of the major search engines allow for targeted searching in minority languages, some of the potential difficulties in using conventional search engines for finding materials in minority languages needs to be addressed.

Now, it must be said that it is not impossible to find material in minority languages using conventional search engines. Entering the correct search terms and possibly using some additional techniques (such as entering the name of a given language in that language or using other limiting terms) can produce results. Such search strategies work better in some languages than others, owing to either the orthographical conventions of that language or other factors. Furthermore, some search engines, e.g. Google, have country specific versions of their database, and thus searches performed for minority language material in that database produce better results. The mere fact that search engines do not provide a mechanism for returning results in a given language does not render the search engine useless. But it does have four effects:

1. the user has to choose another language, invariably the majority language with which that the user is most comfortable, which in turn results in some returns being made to the user in that majority language. Since the user often has some, in many cases significant, skills in that majority language, the user may be tempted to abandon the search in the minority language and follow links in the majority language;
2. it makes it difficult to disambiguate those pages that are properly in a majority language and those that simply contain the key words, but are written in another language;
3. there is a possibility of "homographic interference", that is when two words have the same form, but two different meanings in different languages.

A further effect, though not one faced by the ordinary user, is that it makes it difficult to use search engines to perform linguistic diversity tests. These are particular manifestations of the third problem highlighted by Gerrand – language search restrictions.

Beyond the three problems highlighted by Gerrand above, it must be borne in mind that search engines are primarily commercial constructs that exist to serve a consumer need rather than rigours of academic research. While this fact can be overstated, it nonetheless is a problem that must be addressed and must be proved not be problematic. No one would expect a literature review performed on Amazon.com to be a substitute for a search through university libraries. While Amazon does have some academic material, its purpose is to sell products, and its website prioritises that purpose. Notwithstanding some very thorough scraping of the web performed by the commercial search engines, profit must be assumed to be the primary motive and that this will colour the search and retrieval algorithms and not necessarily completeness to an academic standard. While some good results can be obtained by using search engines, they are not necessarily as reliable as they could or should be.

4. Measuring a Country's TLD

There has been some work done in measuring the 'web' of a particular country, meaning a study of those URLs that match a particular country's URL Country Code, or top level domain ('TLD') (e.g. '.es' for Spain, '.uk' for the United Kingdom). Significant work in this field has been done by Baeza-Yates and Castillo, who have studied the web of Spain (Baeza-Yates, Castillo, & López, 2006) and Chile (Baeza-Yates & Castillo, 2000; Baeza-Yates & Castillo, 2002; Baeza-Yates & Castillo, 2004). It should be noted that the Baeza-Yates studies did specifically address the minority languages of Spain (Basque, Galcian and Catalan). Similar work has been done for the Portuguese (Gomes & Silva., 2003) and Greek (Efthimiadis & Castillo, 2004) TLDs.

The Baeza-Yates and Castillo studies specifically looked for sites that were registered to a particular country's Internet Assigned Numbers Authority (IANA) country code TLD, and also analysed websites that use one of the international generic domains (especially '.com') , which could be verified as being hosted in the target country. Once this was done, they crawled the targeted space and analyse the results.

The 2006 study, took approximately two months to complete and required 45GB of storage space (Baeza-Yates et al., 2006:6). From this dataset it was possible to draw some interesting findings. In terms of language use, the researchers were able to present the following breakdown of languages used in the 'Web of Spain' (Baeza-Yates et al., 2006):

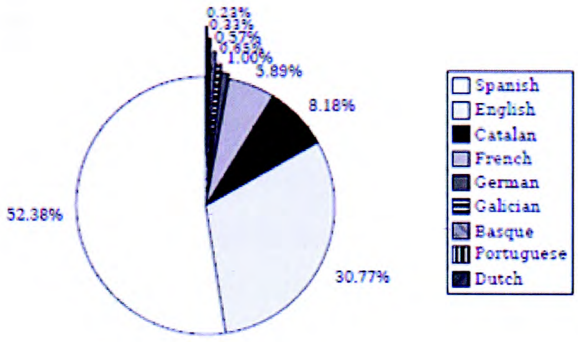


Figure 15 - Web of Spain by Language (2004)

In general, this method is sound and could produce some very interesting results, especially if one is only interested in a particular language or political situation. Two immediate criticisms can be made. Firstly, while it reduces the problem of having to look at the entirety of the web, and only considers a more defined subset, it will eventually run into the same problem as whole-web studies: size and resources. While the 'Web of Spain' may be manageable today, there may come a time when even this subset becomes too large and unwieldy to properly measure. Secondly, it necessarily misses large parts of the web that may be germane to the study. Without a thorough search of the entire internet it is hard to know how much material in a given language exists in other TLDs.

When looking for minority language material, it is important that all the possible sources are identified, since the material available is not as extensive as that for most of the major languages. It is possible that material in any particular minority language would be contained in pages that are not hosted on sites that are not findable by the above methodology. Of course, one could increase the level of the study to include more countries and domains, but at some point the study could easily become too large and too unwieldy. Eventually, the inevitable, logical conclusion is arrived at where the web as a whole needs to be analysed in order to obtain good results, and as the web continues to grow, this problem becomes more serious. The web is not structured linguistically; it is structured politically or by domain. A website where the servers are physically located in the UK may contain information in Welsh, or in any other language. Likewise there is no reason that servers assigned to a .com domain could not contain Welsh material.

Scraping the webspace of a particular country is an interesting idea from the point of view of studying minority languages. But, owing to the structure of internet domain names,

there is simply no direct link between domain name and language. Also, as the web becomes ever larger, it becomes an ever larger problem. Given that it took two months to scrape the Web of Spain in 2006, it would be a very time consuming task to scrape most of the Western European constellations in a similar manner. Finally, the studies did not address the Web 2.0 problem mentioned above. These studies view the web as a static concept, one that was perfectly acceptable in the first decade of the web, but may have problems being adapted to the web going forward.

Therefore, while providing some very interesting results, the study of a particular country's TLD cannot be used to properly measure minority language use on the web.

5. Targeted Studies

For the sake of completeness, reference is made here to a number of small scale studies that look at a number of aspects of the web, mainly with respect to the Welsh language, without attempting to measure any particular aspect. These studies show the range of areas of academic interest

In a review of the literature for the Welsh language, studies have looked at Welsh identity on the web (Thomson & Cunliffe, 2005), internet access in Wales (Richards, 2005), the actual use of Welsh language software by businesses in a part of Wales (Wyn Jones, 2007), party political websites (Cunliffe, 2005), and Welsh on the social networking site Facebook.com (Honeycutt & Cunliffe, 2010). Much work has been done on proper design concepts relating to bilingual websites, with a Welsh focus (Cunliffe, 2004; Cunliffe & Harries, 2005; Cunliffe & Roberts-Young, 2005; Deere & Cunliffe, 2005; Egan, 2000; Harries, 2003; Harries & Cunliffe, 2004; Jarvis, 2000; Welsh Language Board, 2006).

Two interesting studies were commissioned by The Welsh Language Board on websites published by the 22 Unitary Authorities in Wales (Evas, 2001; Gomer, 2003). These provide an interesting look at the state of mandatory Welsh language provision on the web, but significantly were more qualitative than quantitative. That is to say, while they attempted to measure the provision of Welsh on that part of the web, they did so from the point of view of the *Welsh Language Act* and primarily looked at compliance with the main theme of that Act. In particular, no detailed methodology was suggested other than a manual review of these websites. But the body of literature that looks at the holistic situation of Welsh on-line is notably absent.

If we look at similar work that has been done for a number of other languages: Galician and Catalan in Spain (Guinovart, 2003; Mas i Hernàndez, 2003; Romero & Vaquero, 1999); Sardinian in Italy (Mensching, 2000) and minority languages in Africa (Fantognan, 2005; Van Belle, Fellstad, Steele, & van Bakel, 2003), the same tendency can be seen: a focused attempt to study the individual situation of a particular language, but little in the way of general theory.

D. Critique of Prior Methodologies

As detailed above, a number of different methodologies have been employed in the quest to measure language presence on the web. Most studies have suffered from one or more of the following problems:

1. They attempted used a random sampling method that is difficult to continue to use given the ever larger size of the web and the Web 2.0 problem;
2. They used search engines, which are not necessarily reliable from a scientific point of view, notwithstanding that they may produce results that one would expect to see;
3. The studies have not been repeated with enough regularity to permit the observation of long term trends;
4. Most studies have only looked at the world's major languages, and have therefore avoided, for technical or other reasons, the problems inherent in attempting to study minority languages on the web;
5. Most studies give the results for their target languages as relative percentage of the whole web, which is not necessarily useful, given the infinite size and scope of the web, and would not tend to provide useful information for minority language researchers;
6. Some studies tend to concentrate on 'pages per speaker', which do not produce necessarily useful measurements.
7. The principal drawback of these studies is that they have been 'one dimensional': they tend to study only dimension of analysis, typically what is the relative percentage of a particular language within a given set. This does not give necessarily an accurate picture of language use.

E. Wikipedia's Statistics

1. Ranking Language Editions

Wikipedia provides an enormous number of statistics for each language edition of Wikipedia. Wikipedia ranks the various language editions by the number of articles, as is shown in the following screenshot:

All Wikipedias ordered by number of articles [edit]

The languages listed below are Wikipedias which have been created (ordered by number of articles and within each number alphabetically from A to Z). These tables can also be kept up to date by ordinary users. If you feel that an update is needed, simply cut and paste the wiki syntax from this page (which is updated by script fairly regularly) into this page, overwriting the existing tables.

1 000 000+ articles [edit]

№	Language	Language (local)	Wiki	Articles	Total	Edits	Admins	Users	Active Users	Images	Depth
1	English	English	en	3,494,448	22,446,633	430,242,745	1,766	13,536,182	136,263	853,817	564
2	German	Deutsch	de	1,157,244	3,278,567	85,570,744	301	1,119,684	23,767	183,612	88
3	French	Français	fr	1,041,006	4,198,857	63,553,071	150	553,010	16,203	42,710	139

100 000+ articles [edit]

№	Language	Language (local)	Wiki	Articles	Total	Edits	Admins	Users	Active Users	Images	Depth
4	Polish	Polski	pl	754,364	1,361,646	25,432,939	166	401,284	5,343	0	12
5	Italian	Italiano	it	752,034	2,340,506	40,289,159	101	577,318	8,352	82,920	77
6	Japanese	日本語	ja	719,878	1,902,223	35,084,187	63	475,000	11,073	76,691	51
7	Spanish	Español	es	678,595	3,088,035	44,914,135	139	1,670,746	16,307	0	183
8	Dutch	Nederlands	nl	657,292	1,572,883	23,592,038	65	364,600	4,876	20	29
9	Portuguese	Português	pt	645,344	2,443,603	23,269,252	33	806,018	5,678	5,852	74
10	Russian	Русский	ru	630,226	2,231,373	31,453,149	94	581,665	12,217	109,103	91
11	Swedish	Svenska	sv	378,797	1,106,137	13,542,241	101	201,945	3,330	0	45
12	Chinese	中文	zh	335,585	1,163,053	16,288,241	75	926,687	6,402	26,571	80
13	Catalan	Català	ca	296,967	699,275	6,497,196	23	74,535	1,787	6,157	17

Figure 16 - Wikipedia Screenshot - Ranking of editions

As of October 2010, Wikipedia published a list of Wikipedia language editions, ranked by number of articles[15], together with the population figures estimated by Wikipedia.[16] The languages that are the target of this study have been extracted from Wikipedia's list and summarised below:

[15] http://en.wikipedia.org/wiki/Wikipedia:Multilingual_statistics

[16] http://stats.wikimedia.org/EN/Sitemap.htm

Rank	Language	Articles	New (2010)	Popluation	Article/Million
1	English	2,567,509	537,464	1,500,000,000	0.00171
2	German	808,044	161,820	185,000,000	0.00437
3	French	709,312	145,219	200,000,000	0.00355
6	Italian	499,234	144,876	70,000,000	0.00713
9	Spanish	402,430	119,165	500,000,000	0.00080
15	Catalan	133,214	52,866	9,000,000	0.01480
41	Galician	39,627	11,792	4,000,000	0.00991
47	Basque	30,454	9,857	1,000,000	0.03045
56	Breton	21,189	4,772	250,000	0.08476
62	Welsh	18,848	7,220	750,000	0.02513
74	Sicilian	12,809	2,000	8,000,000	0.00160
81	Frisian	9,215	3,701	650,000	0.01418
93	Irish	7,262	1,851	530,000	0.01370
95	Scottish	6,658	2,134	70,000	0.09511
142	Manx	1,880	1,596	2,000	0.94000
148	Cornish	1,563	231	245	6.37959

Table 9 - Wikipedia Ranking of Languages (2010)

If a simple bar chart of these figures is provided, the different orders of magnitude between English and the other languages, both majority and minority, can be readily perceived. These results are not surprising, and graphically illustrate the situation of minority languages today.

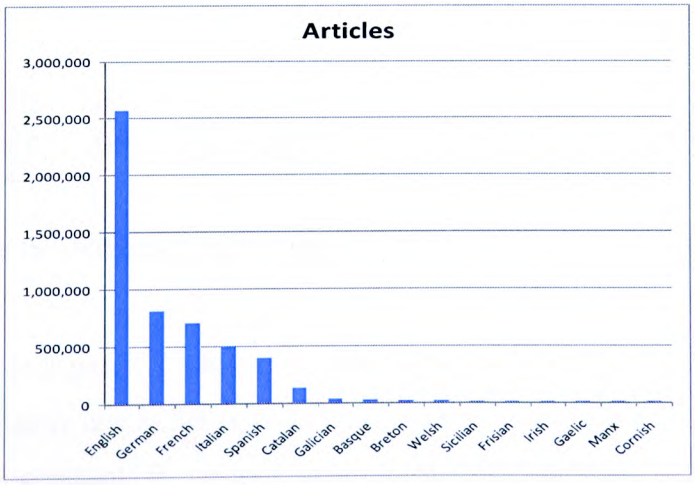


Figure 17 - Wikipedia Language Editions (partial) by Article Number (2008)

The following table shows the extracted data for article counts only, as well as some comparative data, for some of the languages that are the target of this study:

Rank	Language	Articles (2010)	Articles (2008)	Change
1	English	3,494,615	2,567,509	36%
2	German	1,157,244	808,044	43%
3	French	1,041,006	709,312	47%
5	Italian	752,034	499,234	51%
7	Spanish	678,595	402,430	69%
13	Catalan	296,967	133,214	123%
41	Galician	65,095	39,627	64%
45	Basque	61,122	30,454	101%
56	Breton	36,082	21,189	70%
66	Welsh	29,747	18,848	58%
77	Frisian	18,340	9,215	99%
94	Irish	11,759	7,262	62%
104	Gaelic	8,059	2,490	224%
136	Manx	3,644	1,880	94%
156	Sardinian	2,464	844	192%
167	Cornish	2,010	1,563	29%

Table 10 – Wikipedia Article Counts – 2010

The following graphs show the relative size of the major language editions of Wikipedia, based on the data in the above table:

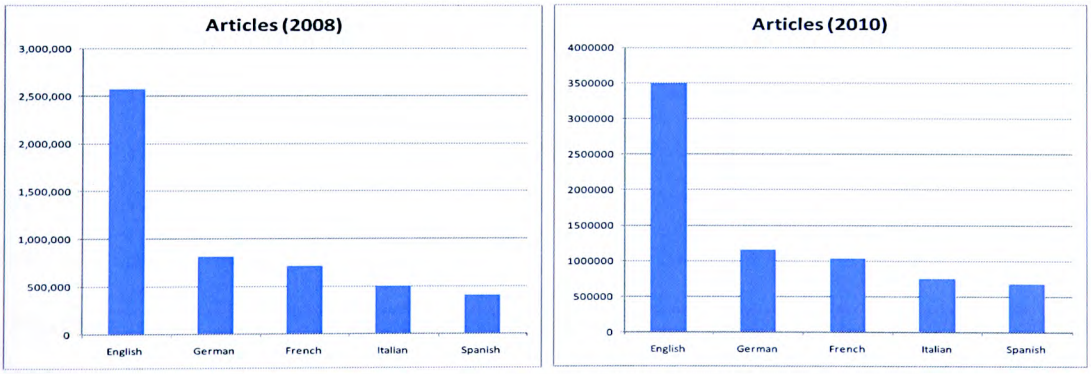


Figure 18 – Wikipedia Major Language Editions Comparison (2008-2010)

As can be seen in Figure 18, while the overall number of articles has risen (in the English edition by approximately one million), the overall relative size of the major language editions has remained fairly constant. If we look at the data from the point of view of number of articles per head of population, we get a similar picture:

1. Wikipedia's Statistics by Population

If we take the above figures and factor for population (using the estimates providing in the above table) we can see a more curious result:

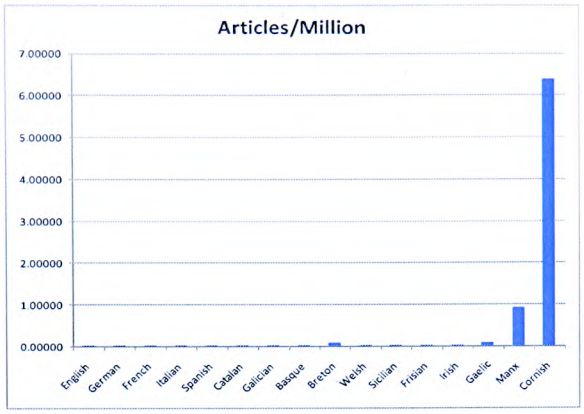


Figure 19 - Wikipedia Language Editions by Population (2008)

As can be seen from the charts above, there are a number of surprises. English is no longer the first ranked Wikipedia edition, but rather three of the other major languages show much stronger results. On the whole, and with the exception of Spanish, the other major European languages outperformed English by a wide margin on this metric. Obviously, there is the fairly bizarre result of Manx and Cornish, two languages that have virtually no native language speakers, but which have an implausible number of pages per speaker. This is a clear demonstration of a known statistical problem whereby small datasets give unreliable results. But even if the results for these two languages are ignored, the problem still remains, as shown in the following chart:

Firstly, it would seem that smaller languages are, pace Sicilian, much more productive on a population basis than the majority languages. Even cursory investigation of minority language material available on the web and on Wikipedia reveals that the content and the range of minority language material is often sparse (as will further demonstrated below). Using population counts, at least when comparing large languages with smaller, lesser-used languages, gives a misleading portrait of web presence

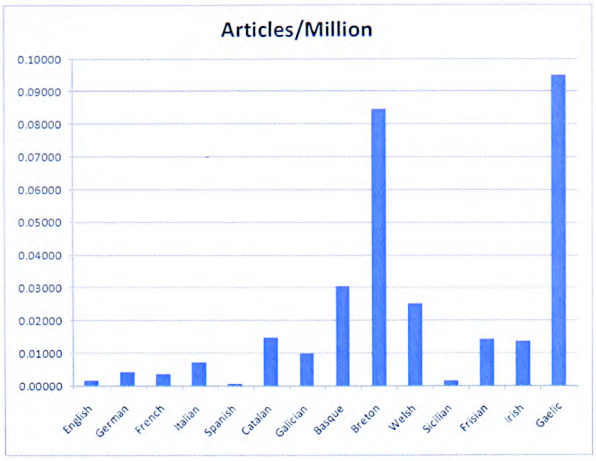


Figure 20 - Wikipedia Language Editions by Population (2008)

.However, if we compare like with like, there are some interesting results. If we compare Irish with Gaelic we have a marked difference between these two comparable language communities. Given that Irish is an official language of the Republic of Ireland we might expect the results to be reversed. However, it cannot be ruled out that there may actually be more activity in the Gaelic language community in Scotland than in the Irish language community. Virtually the same situation exists with respect to Welsh and Breton. These are two closely related languages, but have very different legal statuses in the United Kingdom and France, where the Welsh language has a significantly higher status in the UK than Breton has in the French Republic. It might be thought that this higher status would be reflected here.

F. Other Studies on Measuring Wikipedia

For the sake of completeness a number of studies that were not seen prior to commencement of work on this thesis are summarised briefly here. It would have been ideal to have these studies to hand prior to the development of WkScrape. Since these studies did not inform the methodology, a detailed critique will not be given.

To date there have been a number of preliminary studies that look at Wikipedia and attempt to measure its growth and use. The principal questions that have attracted attention are: measuring the overall size and statistics of particular language editions of Wikipedia (Voß, 2005[a], 2005[b]) and attempts to determine quality from certain mathematical counts and models (Blumenstock, 2008) (Lim, Vuong, Lauw, & Sun, 2006)

One of the first examples of quantitative analyses of Wikipedia is the research work published by Voß in 2005 where he presented some statistics and graphics modelling the evolution over time, and the activity patterns, of the German language edition of Wikipedia. Among other interesting conclusions, he finds that the evolution over time of distinct quantitative indicators, such as the size of the database, the number of articles, the number of active Wikipedians (users who contributed more than 5 times to the project on a given month), the number of very active Wikipedians (users who contributed more than 100 times to the project on a certain month) and the total number of words and internal links, follow an exponential growth rate.

To date the most significant work on measuring Wikipedia is provided by Ortega Soto (Ortega Soto, 2009). He set himself a number of research questions relating to the top ten editions of Wikipedia. In order to test his analyses, the author created a program in Python, WikiXRay, designed to provide quantitative analysis about Wikipedia, generating graphics and providing statistical results for each language edition of Wikipedia. [17]

The program was designed to provide the following types of analysis (Ortega Soto, 2009:55):

- General statistics on the activity of authors, activity per Wikipedia page, and the length and distribution on different namespaces,
- Social structure: an analysis to obtain the best fitted distributions for key descriptive parameters providing information about the distribution of contributions among authors and articles;
- Inequality of the level of contributions from authors and revisions received per article;
- Demography of each language version of Wikipedia;
- Reputation and quality of the articles of each language version of Wikpedia studied;

[17] The current version is hosted at http://developer.berlios.de/projects/wikixray/).

- Evolution: a more in-depth analysis of the evolution in time of the statistical distributions fitted to empirical data in the social structure module, including 3D analysis of the evolution of contributions from the core group of very active users in each month over the remaining history of each language version and a graph of the evolution in time for contributions from logged authors in that language version.

The majority of these questions do not relate *per se* to the question of measuring minority language use on Wikipedia. However, if one were to use the statistics that were provided by WikiXRay from the minority language editions of Wikipedia, there is little doubt that a good picture of minority language use could be provided. Unfortunately, the results of this project were published in 2009, and were not available until the study that is the subject of this dissertation was completed, so it was impossible to use WikiXRay to provide some points of comparison. However, this would provide some interesting future work.

Research work by Vegas et al. presented a method to visualize the evolution over time of the contributions made to a certain Wikipedia article. Many subsequent publications have followed the findings and conclusions presented in this paper to study the working patterns adopted by the Wikipedia community of authors. These authors developed a new software tool, named History Flow, to undertake this analysis

In an article published in 2009 van Dijk directly addressed the problems of measuring minority language editions of Wikipedia (van Dijk, 2009). He notes the phenomenon that Wikipedians seem be obsessed with counting the number of articles in each Wikipedia, and comparing them with other language editions of Wikipedia, this can lead to a race to "increase the number of articles than to write encyclopaedia articles that make their language edition useful to readers" (van Dijk, 2009:236). He notes the phenomena already touched upon where smaller languages' articles in Wikipedia are often much abbreviated and contain little or no real information. In particular he notes the "pseudo-articles" (extremely short, oftentimes one sentence articles that give dictionary like definitions), "geographical stubs" (small articles about towns and villages that give only limited geographical information) and translation type articles, that often provide a minimum of detail, but whose main purpose seems to be to. Thus, relying solely on article counts can give a very misleading impression of the quality and quantity of a particular language edition of Wikipedia.

A small article on the Welsh edition of Wikipedia raises some of the issues that are germane to this study (Jones, 2009). The author lists of number of advantages that a minority language obtains from having a Wikipedia and an active community of Wikipedians: it provides a way to document life in the minority culture; gives writers in a minority language the opportunity to write about topics that are not normally written about in that language; gives momentum to the language and helps enrich the language's lexicon. He notes however, what is obvious to anyone who looks up articles on a minority language edition of Wikipedia: "many articles, it is true, [are] less comprehensive in their scope" (Jones, 2009). This is another way of saying that the presence of Welsh Wikipedia is not that good.

III. Research Methodology

A. Goal of the Study

The narrow goal of this study is to measure the presence of any language on Wikipedia. This will be accomplished by using a program, WkScrape, described in detail below, which randomly samples Wikipedia articles and calculates raw data from those samples. By subjecting the raw data to statistical analysis, a portrait of each language's presence on Wikipedia can be produced and a comparison of such presence with any other language's presence can be made.

In the Introduction, a definition of 'presence' was offered based on a two-dimensional study: measuring both the breadth and depth of material available in a given language. This will be achieved by estimating two different counts of the material contained in a particular language edition of Wikipedia: the number of articles contained in that language edition (the breadth) and an approximation of the amount of information given by the average article (the depth). By using a two-dimensional approach, it is hoped that a more accurate picture can be obtained than by a simple article count, which is the method used by most previous studies of language measurement of the web and is also the means by which Wikipedia measures and ranks the different language editions, as was detailed in Section II.E of the Literature Review, above.

In order to structure the data into groupings that have some real world meaning, two language comparison systems - language constellations and language tier classification - are developed to analyse the different language editions of Wikipedia. The details of these comparison systems are detailed below in Part V – Theoretical Models, below.

B. Research Paradigm

The research paradigm principally employed in this study is Exploratory Data Analysis ('EDA') using multivariate analysis, as originally elaborated by John Tukey (Tukey, 1977). By extracting raw data from random samples drawn from Wikipedia and subjecting those data to EDA statistical analyses an initial picture of the current state of minority language presence in Wikipedia can be drawn. From this initial picture potential hypotheses can be formulated for further study. This methodology makes heavy use of data visualisation techniques in an attempt to see what the data are and what the data can say about the phenomena being studied.

EDA therefore avoids positing *a priori* hypotheses regarding the subject under investigation and instead attempts to create a picture from which future hypotheses can be formed. Since this is the first extensive study that has been made of minority language presence on Wikipedia, there was no prior body of work to form a more concrete methodology, and no initial hypotheses could be formed. While anecdotal evidence and familiarity with the languages involved led to certain expectations, it is very much the goal of this study to let the data speak for itself. It is hoped that this study can draw a number of preliminary conclusions regarding minority language presence on Wikipedia that can form the basis for further work, notably in noting trends in the data and forming working hypotheses that can then be tested.

C. Prior Research Methodologies

As was shown in the Literature Review, previous methodologies have focussed mainly on measuring the number of web pages within a particular set, either from 1) the web as known at the time of the particular study was carried out, or 2) the number of web pages within a search engines database. Using either of these methodologies suffers from the following problems:

1. The impossibly of knowing that the set from which the samples were drawn is actually equal to the universal set. In other words, whether it is certain that the whole of the web was accessible by the methodologies that attempted to randomly sample the entire web or were the contents of the search engines' databases a representative sample of the web?

2. The methodologies ignored the fact that any given webpage can contain varying amounts of information. For example, if it were known that 10% of the webpages within a particular set were in a particular language, would that information alone provide any indication of the depth of information contained on those webpages. It can be readily understood that the mere existence of a set of webpages does not necessarily mean that those webpages contain equivalent information or content as other webpages.

3. Furthermore, as was detailed in the Literature Review, above, prior methodologies, while successful at giving a good picture of majority language material on the web, were not readily useful for minority language studies. This was the reason that other methodologies were sought during the life of this study and why ultimately the scope of the study was narrowed to look at one discrete part of the web.

It is proposed that when measuring presence a breadth and depth approach would produce a more accurate and more complete picture of web presence. In this study, which is using Wikipedia as a test bed, there were a number of data items that could be collected and analysed:

1. The number of articles contained in a Wikipedia edition (the method currently used by Wikipedia);
2. The number of bytes in a Wikipedia language edition;
3. The total amount of informational data in an average Wikipedia article, by looking at the size of a representative article;

However, each of these is only a one dimensional measure. By using Wikipedia's standard methodology of measuring number of articles, only an approximation of the breadth of the contents of any Wikipedia language edition could be estimated. It could be determined that a particular language edition is broader if it contained, say, double the number of articles than another language edition, but there would be no way of knowing how much information is actually contained in that edition. If, on the other hand, the number of bytes is used to calculate the totality of a particular language's database, an estimate could be arrived that would show how much raw information is contained in the database, which would give a good indication of depth, but not of breadth of content. Lastly, an examination of total or average article content would give a good idea of the depth of a given Wikipedia language edition.

From the above, we have one possible measure of breadth: article count, and two possible measures of depth: byte count and article data count. If two of these methods were combined into a composite measure a more accurate picture might be obtained of the state of a particular language's dataset. By measuring both the breadth (in other words, how many 'pages' or other units are available in a given language) and the depth (the total content that is contained in an average unit) and combining them into a composite figure, it is proposed that a more accurate and precise measure of the presence of a particular language edition of Wikipedia can be made.

An examination of a number of Wikipedia articles shows that most articles consist of the following types of data: text; pictures, images, graphs, etc., in a number of formats; footnotes; internal links to other Wikipedia articles; external links. After careful consideration is was felt

that text and footnotes were essentially the same: verbal information. Images obviously constitute, for the most part, a different kind of information, whereas links, internal and external, can constitute quite valuable information, though obviously the text required for a link is quite small. It is believed that external links give, generally speaking, more information since the process of finding and collecting external links is more intensive and potentially more useful to an end-user than internal links, which a user can easily find within Wikipedia itself. Therefore, the decision was made to measure four types of data in a Wikipedia article: the amount of text, the number of images, and the number of external links.

Of course, such a detailed analysis may not be appropriate if other aspects of the web are studied. But the general principle should remain the same. The goal in measuring depth is to understand how much 'content' or 'information' is contained in any measure of presence. With Wikipedia, it is possible to look at text, images and links. Other types of studies may have to look at audio/visual and other types of information.

D. WkScrape

In order to obtain the dataset for a number of language editions of Wikipedia, a specific program - WkScrape – was designed, programmed and tested. The purpose of the program is to scrape and store data from any given language edition of Wikipedia, and then to provide basic data to the researcher in the form of user and machine readable tables containing raw statistical information.

The program first scrapes the index page of the designated language edition of Wikipedia and stores a list of the articles of that language edition is a database. Secondly, the program selects, using a random number generator (to a maximum value set by the researcher), a number of articles from the list of links obtained from the first scrape. Thirdly, a second scrape of Wikipedia is undertaken, in which the contents of each Wikipedia article, drawn from the set obtained in the second step, and a number of raw calculations are performed. The results of the second scrape are stored in a set of machine readable files.

The software has been designed so that virtually any language edition of Wikipedia can be scraped with only minor adjustments needed. To date, the program has been tested on

approximately twenty language editions.[18] All the language editions tested so far have used the Roman alphabet, but there is no reason to believe that the program could not be successfully used to analyse other alphabetic systems, and, using the correct ISO encodings, non alphabetic scripts.

Initial program development took approximately four months to reach an acceptable state whereby it was able to scrape, store and perform basic calculations on Wikipedia articles in two languages: English and Welsh. The first beta version of the program was hard-coded to only scrape only English and Welsh, using two different methods for each language. Development continued for approximately another four months, removing the inevitable bugs and errors, and scaling the program so that it could work with any language. The structure of the program was later streamlined, necessitating a major rewrite of the program.

The design of the program needed to take account of the following factors:

1. All programming tools and libraries needed to be open-sourced (in the sense that they could be used and implemented without payment of licence fees or other limitations). Open source also allowed for fast deployment of the various libraries. This has the added benefit that there are no restrictions on any eventual distribution of the program;
2. Ideally, the program should be able to be run on a number of operating systems. To date the program has been successfully run on Windows (XP, Vista and Windows 7) and Linux (tested on several distributions, including Ubuntu);
3. The program needs to be able to scrape and test a large number of languages, and, ideally, all the language editions of Wikipedia.

It is believed that all the design considerations were met. WkScrape was written in Java 1.5, using the Eclipse Galileo IDE (version 3.5). The program made use of the following libraries, and any dependent libraries required:

---

[18] Besides English and Welsh, the following languages have been successfully tested: Breton (BR), Catalan (CA), Danish (DA), German (DE), Spanish (ES), French (FR), North Frisian (FRR), West Frisian (FY), Irish (GA), Gaelic (GD), Galician (GL), Manx (GV), Interlingua (IA), Icelandic (IS), Italian (IT), Cornish (KW), Latin (LA), Maori (MI), Dutch (NL), Norfolk (PIH), Sardinian (SC) and Saterland Frisian (STQ).

- HTMLParser, version 1.6[19], a library containing a number of classes useful for web scraping, link extraction and other tools necessary for getting and manipulating web pages;
- Apache Lucene 1.4.3[20], an information retrieval package, used to store the data acquired for article analysis;
- Dom4j, version 1.6.1[21], a library used for working with XML files, which is the principal method for storage of some of the program's data
- Apache log4j, version 1.2.15[22], a library of logging software, necessary to keep track of the numerous data files produced by the program.

All of the above libraries were open source and were used in accordance with their licence agreements. The program has ten classes as is shown in detail below.

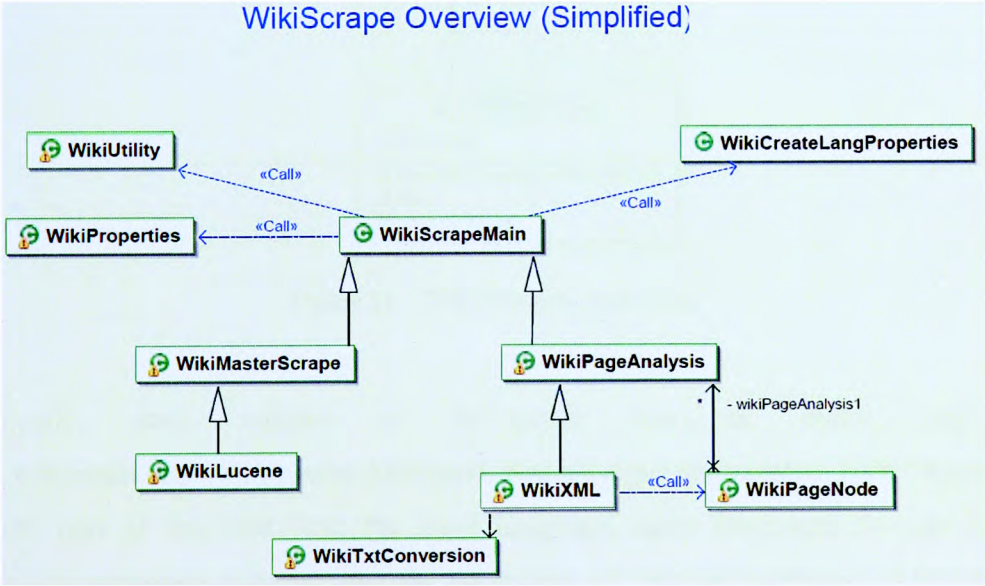


Figure 21 - WkScrape Overview

[19] http://htmlparser.sourceforge.net/

[20] http://lucene.apache.org/java/docs/index.html

[21] http://dom4j.sourceforge.net/

[22] http://logging.apache.org/log4j/1.2/

The program has two high level functions: (1) extract from each language edition's index a list of links to all the articles contained in each language edition of Wikipedia, which is handled by the class WikiMasterScrape, and (2) analyse a sample set of articles contained in the particular language edition in order to measure a number of features of each Wikipedia article, chiefly, the number of words, the number of images and the number of links contained in the article. This last analysis is handled by WikiPageAnalysis.

1. Class WikiMasterScrape

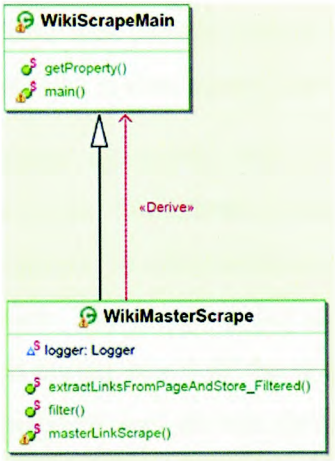


Figure 22 – WikiMasterScrape Class

Typically, each edition of Wikipedia has an index page (e.g. http://en.wikipedia.org/wiki/Special:AllPages). Some languages translate both "Special" and "AllPages" part of the URL into the local language, some languages do not. For each language it is necessary to know how the equivalent of "Special:AllPages" in that language (e.g. CY – "Arbenning:AllPages"; GA – "Speisialta:AllPages"; DE – "Spezial:Alle_Seiten"). Each edition of Wikipedia has to be examined to determine how this translation is handled. In order to correctly scrape a particular language edition, the program needs eight further pieces of information. These are entered by the user in a text file that is converted by the program into a .properties file. The information needed is not onerous and can be obtained by the user within a few minutes.

```
# Cymraeg (CY) language configuration file
# language code is a two letter abbreviation
[language code]: cy
# majority language code is a two letter abbreviation
[majority language code]: en
#the allpages index url
[all pages index url]: not known at this time
# the seeding url (the first link on the all pages url)
[seeding url]: http://cy.wikipedia.org/w/index.php?title=Arbennig:AllPages&from=%22Bugs%22+Moran
# extract the next page string from 'previous page' to 'next page' (e.g. Precedente pagina .*Tote_le.*from.*Sequente pagina)
[next page link string]: bloc cynt gan ddechrau .*AllPages.*from.*Y bloc nesaf
# an alternate next page if necessary (put 'AD Note: not used')
[next page link string alternate]: AD Note: not used
# the string for 'Special:AllPages' in the particular language
# will become - /w/index\\.php\\?title=Special:AllPages&amp;from=[\\p{Graph}\\p{L}]+
[next page url str]: Arbennig:AllPages
# a semi-colon delimited list of the hallmark strings for special pages, etc
[excluded str short]: categori;arbennig;wicipedia;hafan;materion_cyfoes
```

Figure 23 – WkScrape Language Configuration File

Once the index page is identified, it is then possible to get scrape all the article URLs that are contained in that index for that edition of Wikipedia. The program scrapes each link, and tests for validity, rejecting certain links that contain certain keywords, notably those that do not refer to actual articles. The list of links is then stored in a Lucene database.

There are a number of problems associated with this method of obtaining links. Wikipedia indices are not always accurate. Sometimes articles are listed in the index but are no longer existent or are actually 'referers' to other articles, meaning that link does not point to its own article, but rather to another. The former is not as problematic as the latter. Links that are no longer existent are stored during step 1 of the process, but are rejected at stage 2, and therefore are ultimately not analysed and do not form part of the eventual analysis. However, referrer articles are not rejected at state 2 (since they appear to be valid articles). This will result in an article being sampled and counted twice. For example, in the Welsh language edition of Wikipedia, the index lists separate links for 'Cymraeg', and the referrrers 'Gymraeg', 'Chymraeg' and 'Nghymraeg' (which are all mutated forms of the same radical – 'Cymraeg'). The three referrers refer to the same article. Any random count that includes any one of these referrers will be counted as if they were separate and independent articles. At the moment, no satisfactory method has been arrived at for solving this problem, though a method could be added to the program to check for this problem, and would be the subject of future work.

A full scrape of the English language edition of Wikipedia, the largest language edition, with about eight million articles listed in the index, took about three hours and forty minutes on a low quality machine with a Celeron processor and 2GB of memory. The smaller language editions take only a few minutes. The processing power needed is not large and the

principal delays in performance are related more to the available internet bandwidth of the connection and the ability of Wikipedia's servers to respond to the scrapes.

2. WikiPageAnalysis

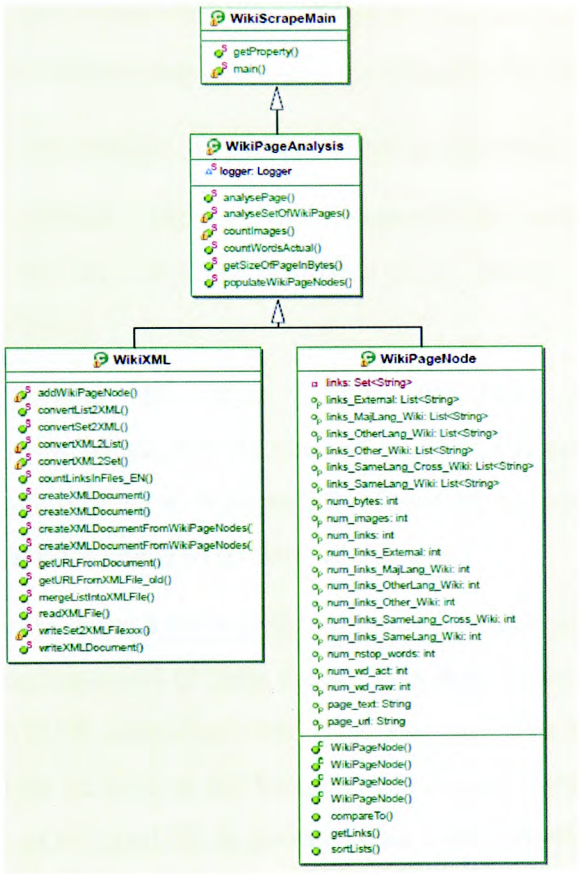


Figure 24 - WikiPageAnalysis Class Diagram

The WikiPageAnalysis class is the class that handles the analysis of sampling of articles that are randomly drawn from the total population of the links obtained from stage 1 scrape of the index pages. The sample rate is determined by the user and can be any arbitrary number between 1 and the maximum number of articles contained in the edition's index, obtained during the stage 1 scrape.

WikiPageAnalysis uses two classes to complete the task. For each article in the sample, an instance of WikiPageNode is created. WikiPageNode counts the items of interest for each Wikipedia article:

- Num_bytes – the raw number of bytes for the article;
- Num_wd_raw – the raw number of words contained in the article;
- Num_wd_act – the actual number of words that form content of the article, i.e. words contained in the article, excluding the words that are used by Wikipedia as background information and formatting text;
- Num_images – the number of images contained in the article;
- Num_links – the number of links contained in the article;
- Num_links_external – the number of external links contained in the article, i.e. those links that do not cross reference other articles in the same or another Wikipedia edition;

In addition, a number of other counts of links were attempted (e.g. num_links_MajorLang_wiki, which was an attempt to count the number of links that were made to a certain language edition of Wikipedia.) However, the results for these tests were not successful and further work needs to be done in this area.

Finally, WikiPageNode stores a copy of the complete contents of the article and creates a number of lists for the various types of links found. This data, along with the various counts, are stored, using the WikiXML class, into .xml files. The .xml files contain a snapshot of the sample, and could possibly be used in the future to run further tests on the same sample. A summary of the contents of the .xml file is also saved as a tab delimited .txt file, which can be easily imported into data processing programs such as SPSS or Microsoft Excel.

It can be noted here that many articles contain footnotes, which do, of course, add additional information. Footnotes were not counted separately since the actual verbal content of the footnote will be picked up by the method that calculates the variable num_words_actual, and the links contained in the footnotes will be counted by the various link counting and storing methods of the program.

E. Testing the Data

1. Validity of the data

a) Reliability of WkScrape's Database

A proper statistical analysis requires a proof that the sample results are valid and reliable. In this case, it would have to be shown that WkScrape performs an appropriate scrape of Wikipedia articles and that the various statistics produced by the program are an accurate representation of Wikipedia as a whole.

According to Wikipedia's home page[23], on 19 February 2009, the English edition of Wikipedia contained 2,748,005 articles. WkScrape produced an index of 6,113,189 pages. Thus, over twice the number of pages was obtained from a full scrape of Wikipedia's index page than Wikipedia claims for its total article count. For the scrape of the Welsh edition of Wikipedia, an opposite result was obtained. The Welsh edition's main page[24] claims 21,446 articles, whereas WkScrape's database stored exactly 18,000 articles.

There is therefore an initial validity problem in that in the scrape of the English edition produces almost double the expected number of results for the English edition, while for the Welsh edition, slightly less than the number of expected results was obtained. There are a number of possibilities for this problem: (1) WkScrape is not accurately scraping the index pages of the English and Welsh language editions of Wikipedia; (2) the index from which WkScrape is obtaining the data is inaccurate or not up to date; or (3) the article count given by Wikipedia is not accurate.

An examination of the results of a preliminary test (see below), where 50 samples from both the English and Welsh editions of Wikipedia were drawn showed that there were three invalid cases from the English Wikipedia and no invalid cases from the Welsh Wikipedia. The three invalid cases from the English sample were to two 'disambiguation pages', which are pages that contain links to articles that have similar names but refer to different concepts. The third invalid case was to an article which stated 'Wikipedia does not have article with this exact name'. Wikipedia mentions a number of possibilities for receiving this result including

---

[23] http://en.wikipedia.org/wiki/Main_Page

[24] http://cy.wikipedia.org/wiki/Hafan

that there may be a problem in updating the database, a problem with case sensitivity, or that the page was deleted. Upon a direct verification of the case it was discovered that there was a similar article, but with a slightly different URL string. Thus, it would appear that the Wikipedia index database was not updated, and that this accounted for the error.

Therefore, in the English language edition's index includes disambiguation articles and improperly referenced articles, and it can be concluded that the difficulty arises within the Wikipedia system and not with WkScrape. Simply put, there are a lot more Wikipedia pages indexed than there are articles. Since WkScrape scrapes the Wikipedia index for links, rather than articles, it can be concluded that WkScrape is producing a valid scrape of the contents of the Wikipedia index. In the event that the index contains references to invalid cases then this can be dealt with at the data analysis stage.

Obviously, a search for hallmarks and specific patterns in the links contained in the index would constitute the most obvious method of eliminating such invalid entries, but care needs to be taken since hallmark text can often appear in otherwise valid articles. Without such certainty that these hallmarks do, in fact, refer to invalid articles, it is impossible to do any further improvements to the WikiMasterScrape class. In this case, it may be more prudent to have a larger data set than required and then to remove invalid entries afterwards.

b) Completeness of WkScrape's Database

A second test of WkScrape's accuracy with respect to obtaining a complete scrape of the index is whether it contains all the articles that do currently exist within the Wikipedia edition's database. Since we cannot tell from an examination of the index what it does not contain, a method that randomly selects articles from the Wikipedia database must be used and then verification made as to whether such an article appears in the WkScrape database.

Wikipedia contains a "Random article" feature that allows the user to obtain an article selected randomly by Wikipedia's software. No verification has been made as to the algorithm used by Wikipedia to produce the random article, but it appears to perform as required in that it returns a seemingly randomly generated article. A small test was performed for fifty such articles in the English language edition of Wikipedia:

Number	Article obtained at Random	Database #	Found
1	http://en.wikipedia.org/wiki/The_Strasbourg_Conference	5396369	yes
2	http://en.wikipedia.org/wiki/Sergei_Bondarchuk	4847870	yes
3	http://en.wikipedia.org/wiki/Gerhard_Sandbichler	2072685	yes
4	http://en.wikipedia.org/wiki/Daniel_J._Kim	1367949	yes
5	http://en.wikipedia.org/wiki/Kimberley_Girls%27_High_School	2933098	yes
6	http://en.wikipedia.org/wiki/Anais_Catala	372496	yes
7	http://en.wikipedia.org/wiki/Parmelia Note: this is a disambiguation article	4122853	yes
8	http://en.wikipedia.org/wiki/Sin_Na-Hee	4928869	yes
9	http://en.wikipedia.org/wiki/William_Edmond_Armitage	5925712	yes
10	http://en.wikipedia.org/wiki/Thematic_structure	5422155	yes

Table 11 – 10 of the 50 Randomly Selected Articles from Wikipedia

Using Luke, a Lucene database explorer, it was then possible to determine whether the article was entered into the database by the WikiMasterScrape class. In all fifty cases, the articles were contained in the database. The same test was performed for the Welsh edition of WkScrape's database, with the same results obtained, namely that 50 random articles were obtained by requesting a random article.

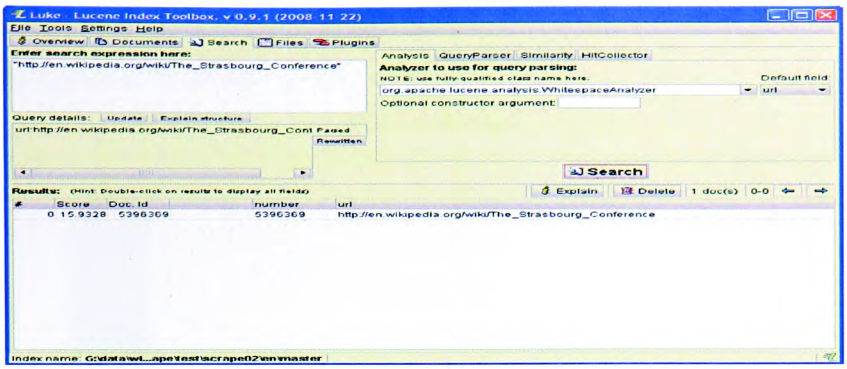


Figure 25 - Verifying the existence of an article in the DB

c) Conclusion regarding validity of WkScrape

On the basis of the above tests, it can be concluded that WkScrape does what it is required to do in stage 1: it scrapes the entirety of the index of a given language edition of Wikipedia and stores all the links contained in that index in WkScrapes's database. If there is a problem, it is that it tends to obtain more links to articles than there are articles currently contained in the Wikipedia database, but that problem seems to be related to Wikipedia rather than the program itself. As such, further work needs to be done to see whether the problem can be reduced or mitigated.

2. Analysis of the Data – English and Welsh

In order to give a foretaste of what is to come, the following is a brief view of some of the salient statistics that are produced by the WkScrape program. The following is based on a small test scrape and an analysis performed on 9 April 2009 on both the English and the Welsh editions of Wikipedia. The sample size was deliberately set to a small number, 100 samples, simply to highlight the scope of the problem. The relevant averages obtained for num_words_raw are summarized in the following table produced when the summary of results file, produced by WkScrape, was imported into SPSS and basic statistical analysis performed:

	Mean	Median	Mode	Std. Dev.
EN	1751.89	782	233	2652.7
CY	313.89	249	125	225.3
	Minimum	**Maximum**	**Range**	**Sum**
EN	208	15359	15151	169933
CY	89	1482	1393	30447

Table 12 - Comparison of statistics EN-CY (100 Samples)

Which can be graphed as follows:

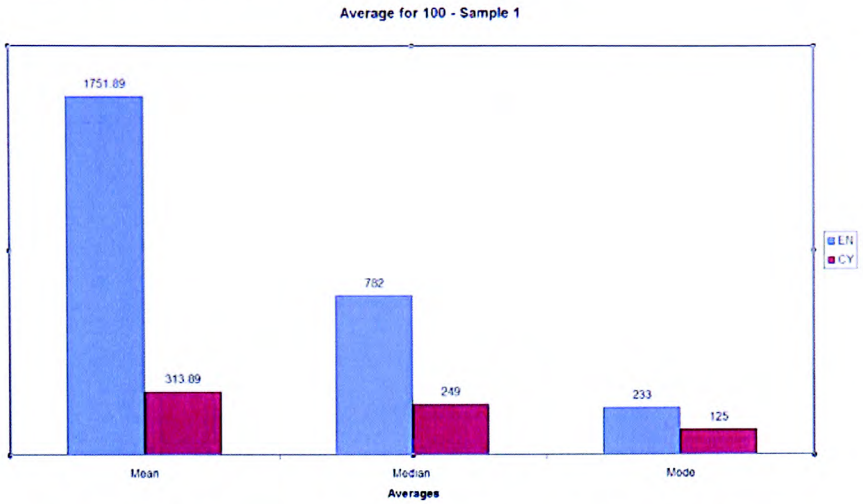


Figure 26 - Comparison of statistics num_words EN-CY (100 Samples)

There is obviously a very large difference between the mean number of words for the English edition of Wikipedia when compared to the Welsh edition. This will have a significant impact on the overall calculation of the presence of the Welsh versus the English editions of Wikipedia. Not only does the English language edition contain many more articles, but the average number of words per article is significantly larger.

The following charts show the frequencies for the two language editions, sorted by number of words per article (please note that they are not on the same scale, as indicated):

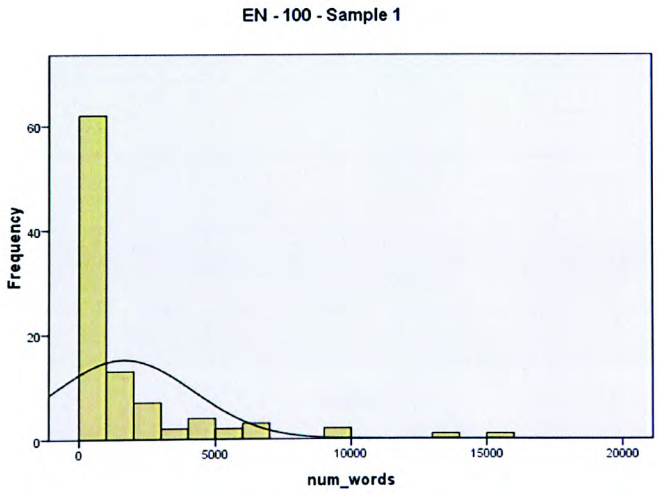


Figure 27 - EN - 100 Samples Distribution

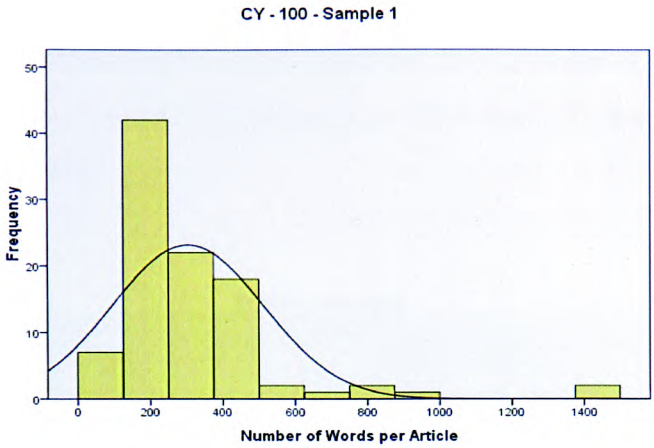


Figure 28 - CY -100 Samples Distribution

This data is even more clearly seen in the following box plot. As in all box plots, the top of the box represents the 75th percentile, the bottom of the box represents the 25th percentile, and the line in the middle represents the 50th percentile. The whiskers (the lines that extend out the top and bottom of the box) represent the highest and lowest values that are not outliers or extreme values. Outliers (values that are between 1.5 and 3 times the interquartile range) and extreme values (values that are more than 3 times the interquartile range) are represented by circles beyond the whiskers.

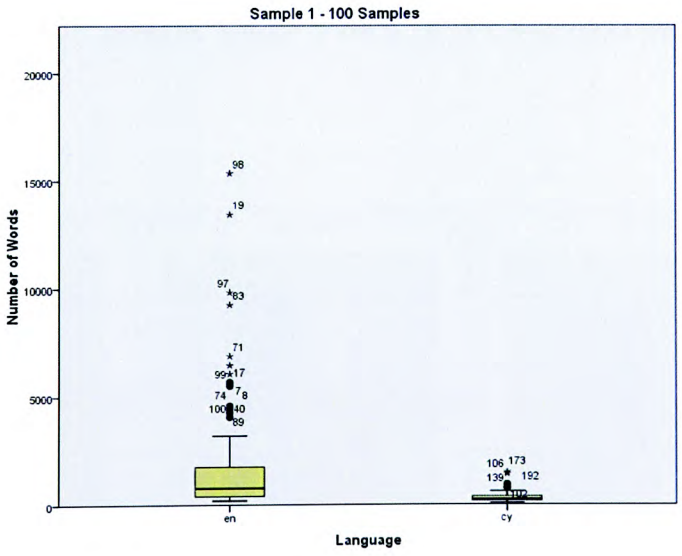


Figure 29 - Box plot of EN-CY (100 Samples)

While the above box plot shows clearly the difference in the ranges between the English and Welsh samples, the following box plot, which has been cleaned of outliers, shows more clearly that the average number of word per article differs markedly between the English and Welsh editions of Wikipedia.

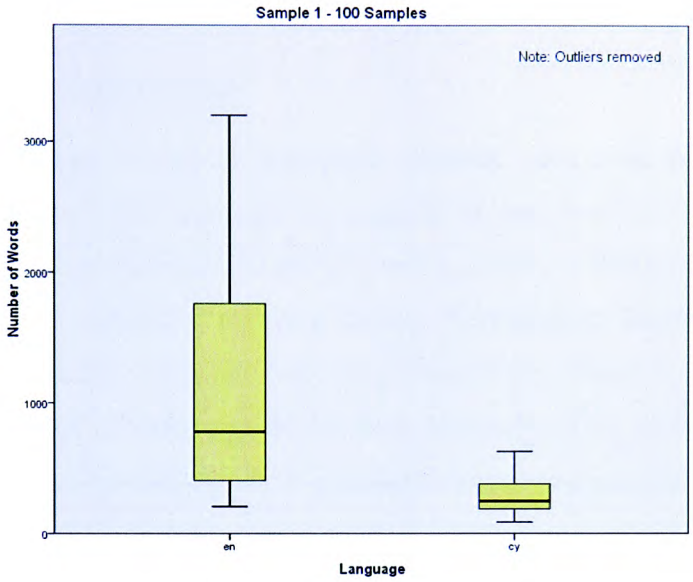


Figure 30 - Box plot of EN-CY (100 Samples) – Detail

Likewise, the same differences appear if we look at the average number of images and links.

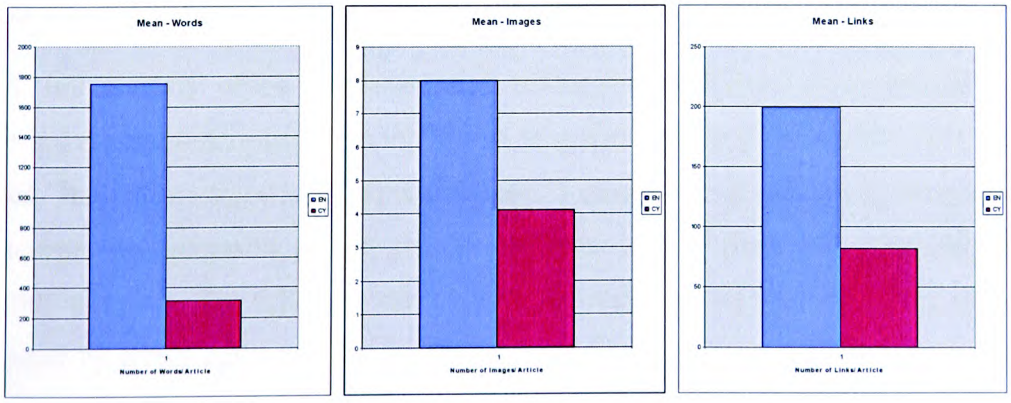


Figure 31 - Means - Words, Images, Links EN-CY (100 Samples)

This small test of only 100 samples gives a good idea of the differences in depth of coverage between the English and Welsh editions of Wikipedia. As was shown in the graphs and table above, the English language edition's average article is much large and contains more words, images and links than its Welsh equivalent. Thus, the Welsh edition is not only narrower in terms of breadth of coverage, but is shallower in terms of depth of coverage.

3. Choosing the sampling size

The English language edition of Wikipedia contains more than three million articles. While it would be conceivable to scrape the entirety of that database, it would be a rather large task and, most importantly, it would put more strain on Wikipedia's servers, not to mention requiring more significant memory storage. Furthermore, WkScrape is still in beta stage, and improvements are still being made. In a future study, however, it may be feasible to consider a full scrape of all language editions of Wikipedia. This would avoid any errors produced by sampling and would provide a complete and accurate picture.

To date, WkScrape has been tested on sample sizes of 3 to 10,000 articles. In choosing the optimal sample size, we would normally use the largest size possible. However, in practice, as the magnitude of the sample size increases, so does the overhead in terms of time needed to scrape the sample and the amount of memory needed. In order to correctly determine an acceptable sample size, it was decided to scrape the entirety of the Welsh language edition of Wikipedia. By doing so, an accurate state of the contents of the Welsh language edition of Wikipedia could be obtained, and from that basis, a good idea of the ideal sample size could be determined.

A 'full analysis' of the Welsh language edition was performed on 2 April 2009. A scrape of the index page determined that the Welsh language edition of Wikipedia contained 38,736 articles. The entire scrape took approximately 28 hours and 15 minutes to complete. In order to determine an acceptable sample size, six different sample sizes: 100, 250, 500, 1000, 2000 and 3000 were run. The following tables show the results of the analyses of the six samples as processed by SPSS against a full analysis of the Welsh language edition, containing 38736 articles, as well as samples in the range of 3000 to 100 samples:

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	38736	38736	38736	38736	38736	38736
	Missing	0	0	0	0	0	0
Mean		25727.78	412.79	191.61	1.09	110.10	2.60
Median		22812.00	321.00	100.00	1.00	90.00	2.00
Mode		18206	275	0	0	50	2
Std. Deviation		12787.039	379.487	364.313	4.079	74.726	2.378
Variance		1.635E8	144010.407	132723.725	16.635	5584.018	5.655
Range		678185	10829	10801	289	3489	121
Minimum		4872	110	0	0	30	2
Maximum		683057	10939	10801	289	3519	123
Sum		996591187	15990027	7422181	42095	4264900	100577

Table 13 - Statistics - Full Scrape CY (38736 Articles)

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	3000	3000	3000	3000	3000	3000
	Missing	0	0	0	0	0	0
Mean		25877.69	411.92	189.87	1.06	110.93	2.54
Median		23010.00	324.00	99.00	1.00	91.00	2.00
Mode		18745	304	0	0	54	2
Std. Deviation		12473.954	398.364	394.942	2.712	72.330	2.018
Variance		1.556E8	158693.577	155978.797	7.354	5231.678	4.073
Range		202176	10796	10801	57	714	64
Minimum		14289	143	0	0	32	2
Maximum		216465	10939	10801	57	746	66
Sum		77633059	1235760	569599	3168	332789	7634

Table 14 - Statistics - Scrape CY (3000 Articles)

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	2000	2000	2000	2000	2000	2000
	Missing	0	0	0	0	0	0
Mean		25730.40	417.25	200.63	1.06	110.43	2.58
Median		23136.00	320.00	104.00	1.00	91.00	2.00
Mode		16667[a]	288	0	0	54	2
Std. Deviation		11652.301	428.533	427.717	2.930	71.056	2.195
Variance		1.358E8	183640.394	182942.217	8.586	5048.979	4.818
Range		148892	10828	10801	70	716	64
Minimum		13608	111	0	0	30	2
Maximum		162500	10939	10801	70	746	66
Sum		51460804	834495	401265	2123	220861	5168

a. Multiple modes exist. The smallest value is shown

Table 15 - Statistics - Scrape CY (2000 Articles)

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	1000	1000	1000	1000	1000	1000
	Missing	0	0	0	0	0	0
Mean		25482.62	405.48	188.84	1.03	109.82	2.56
Median		22449.50	317.50	97.00	1.00	86.00	2.00
Mode		16243[a]	221	0	0	76	2
Std. Deviation		11548.776	369.992	367.677	2.273	77.327	2.810
Variance		1.334E8	136893.811	135186.702	5.166	5979.517	7.898
Range		97385	6443	6516	37	661	78
Minimum		13608	111	0	0	30	2
Maximum		110993	6554	6516	37	691	80
Sum		25482618	405481	188838	1025	109824	2559

a. Multiple modes exist. The smallest value is shown

Table 16 - Statistics - Scrape CY (1000 Articles)

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	500	500	500	500	500	500
	Missing	0	0	0	0	0	0
Mean		24730.09	378.99	154.00	1.01	105.18	2.48
Median		22358.00	305.00	94.00	1.00	87.00	2.00
Mode		18103[a]	267	0	0	49	2
Std. Deviation		8998.793	212.818	193.069	2.055	62.318	1.413
Variance		8.098E7	45291.467	37275.683	4.222	3883.509	1.998
Range		61427	1583	1514	31	316	15
Minimum		14875	141	0	0	36	2
Maximum		76302	1724	1514	31	352	17
Sum		12365044	189494	76998	507	52589	1239

a. Multiple modes exist. The smallest value is shown

Table 17 - Statistics - Scrape CY (500 Articles)

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	250	250	250	250	250	250
	Missing	0	0	0	0	0	0
Mean		26061.95	404.65	180.12	.93	113.38	2.54
Median		22784.00	324.00	105.50	1.00	88.50	2.00
Mode		16236[a]	286	0	0	81	2
Std. Deviation		13648.833	291.223	257.853	1.614	81.965	1.397
Variance		1.863E8	84810.936	66488.186	2.605	6718.302	1.953
Range		127166	2790	2810	17	621	14
Minimum		15085	148	0	0	37	2
Maximum		142251	2938	2810	17	658	16
Sum		6515487	101162	45030	232	28346	634

a. Multiple modes exist. The smallest value is shown

Table 18 - Statistics - Scrape CY (250 Articles)

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	100	100	100	100	100	100
	Missing	0	0	0	0	0	0
Mean		23075.38	333.45	121.78	.70	94.89	2.27
Median		21406.50	287.00	81.50	1.00	83.50	2.00
Mode		15007[a]	263[a]	0	0	47[a]	2
Std. Deviation		6299.542	153.750	139.879	.905	44.882	.664
Variance		3.968E7	23638.937	19566.153	.818	2014.422	.442
Range		29364	915	615	5	198	3
Minimum		15007	160	0	0	35	2
Maximum		44371	1075	615	5	233	5
Sum		2307538	33345	12178	70	9489	227

a. Multiple modes exist. The smallest value is shown

Table 19 - Statistics - Scrape CY (100 Articles)

The following table shows the averages (mean, median and mode) for the different sample rates, as well as the percentage of the difference between each of the averages for each sample from the particular average of the full set.

	Mean		Median		Mode	
Full Scrape	25727		22812		18206	
Sample	Reported	Percent of Difference	Reported	Percent of Difference	Reported	Percent of Difference
3000	25877	0.58	23010	0.86	18745	2.88
2000	25730	0.01	23136	1.40	16667	-9.23
1000	25482	-0.96	22449	-1.62	16243	-12.09
500	24730	-4.03	22358	-2.03	18103	-0.57
250	26061	1.28	22784	-0.12	16236	-12.13
100	23075	-11.49	21406	-6.57	15007	-21.32

Table 20 - Percentage of Difference from Full Scrape

The mean for the 3000 sample rate only differed from the true mean of the full scrape by only 150 bytes or by .58%. The 100 sample rate showed a much wider degree of difference (11.49%, in this case be several hundred bytes less), which is deemed to be significant. However, as can be seen, there is a wide range of differences. In particular, the 2000 sample rate results had only a very slight divergence from the true mean. But, if we look at the other averages it can be seen that, overall, the 3000 sample rate showed the least divergence from the actual averages. This can be best demonstrated by the followed chart, which shows the

cumulative differences from the true averages, with the absolute value of the averages combined to show a combined total of differences for the three measures of average.

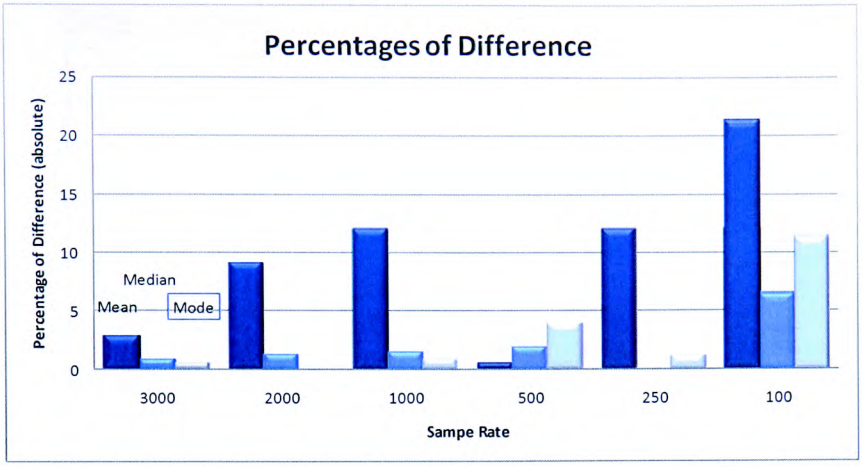


Table 21 - Percentages of Difference for Samples

As would be expected, the largest sample size of 3000 showed the least difference from the various averages, with a general increase in difference as a smaller sample rate was selected. Clearly a sample rate of 100 is too small, since, in this case an average for this sample rate can vary by more than 20% (as in the case of the mode) from the true average. By selecting a sample size of 3000 articles, there is a margin of error of approximately 2.88%. Thus all results hereinafter after a margin of error of +/- 3%.

While it may seem an obvious point that the highest of the chosen sample rates would produce the best results, it was nevertheless a useful exercise to demonstrate this by reference to the true means that was produced by the analysis of the universal set. Obviously a higher sampling rate would be even better and would produce a lower margin of error. However, it was decided to stick with a sample rate of 3000 articles as this was more efficient in terms of time and resources, easier on Wikipedia's servers and the margin of error was deemed to be acceptable.

F. Population Estimates

As was mentioned above, Wikipedia uses its own estimates to rank Wikipedia editions by population. The data provided by Wikipedia needs to be looked at closely. The following

table gives, for 2010, the count of articles as well as an estimate of the population of speakers for each of the target languages of this study:

Rank	Language	Articles	Popluation	Art/Pop (2010)
1	English	3,494,615	1,500,000,000	0.00233
2	German	1,157,244	185,000,000	0.00626
3	French	1,041,006	200,000,000	0.00521
6	Italian	752,034	70,000,000	0.01074
9	Spanish	678,595	500,000,000	0.00136
15	Catalan	296,967	9,000,000	0.03300
41	Galician	65,095	4,000,000	0.01627
47	Basque	61,122	1,000,000	0.06112
56	Breton	36,082	250,000	0.14433
62	Welsh	29,747	750,000	0.03966
81	Frisian	18,340	650,000	0.02822
93	Irish	11,759	530,000	0.02219
95	Gaelic	8,059	70,000	0.11513
142	Manx	3,644	2,000	1.82200
148	Cornish	2,010	245	8.20408
159	Sardinian	2,464	2,000,000	0.00123

Table 22 - Wikipedia Data for Target Languages (2010)

Some comments need to be made with respect to Wikipedia's population estimates. Whilst it is not impossible that one quarter of the world's population has some knowledge of English, the figure of 1,500,000 for English is probably too large. Given that the US has about 300 million first language speakers, with the UK adding another 60, with a further 60 million in Canada, Australia and New Zealand, it is difficult to see how there are 1.5 billion native languages speakers of English. Likewise the figures for German and French are probably overstated as well. It is very difficult to estimate language populations, since in many cases reliable data are simply not present. Two other organisations that measure language have also provided some estimates, and it is useful to compare them. The following table gives the figures obtained from Omniglot.com and from Ethnologue.com:

Language	Wikipedia	Omniglot	Ethnologue	Mean
English (EN)	1,500,000,000	608,000,000	328,000,000	812,000,000
German (DE)	185,000,000	121,000,000	90,300,000	132,100,000
French (FR)	200,000,000	265,000,000	67,800,000	177,600,000
Italian (IT)	70,000,000	60,000,000	61,700,000	63,900,000
Spanish (ES)	500,000,000	417,000,000	329,000,000	415,333,333
Welsh (CY)	750,000	797,717	537,870	695,196
Irish (GA)	530,000	353,000	391,470	424,823
Gaelic (GD)	70,000	60,000	66,780	65,593
Manx (GV)	2,000	1,689	300	1,330
Cornish (KW)	245	300	100	215
Breton (BR)	250,000	365,000	500,045	371,682
Dutch (NL)	27,000,000	20,000,000	21,730,290	22,910,097
W. Frisian (FY)	650,000	450,000	467,000	522,333
N. Frisian (FRR)	10,000	8,000	10,000	9,333
Sa. Frisian (STQ)	2,000	2,000	5,000	3,000
Danish (DA)	6,000,000	5,500,000	5,450,000	5,650,000
Icelandic (IS)	320,000	300,000	230,000	283,333
Catalan (CA)	9,000,000	12,000,000	11,530,160	10,843,387
Galician (GL)	4,000,000	3,000,000	3,185,000	3,395,000
Sicilian (SC)	2,000,000	1,200,000	1,045,000	1,415,000

Table 23 - Estimate of World Language Populations

It is interesting to note that that, with a few exceptions, the Ethnologue figures are quite low. The combined populations of the United States and the United Kingdom alone exceed the figure given for English. While it is true that the USA has a large Hispanic population, and the major English-speaking nations all have sizeable immigrant populations, as estimate of only 328,000,000 for English is a very conservative figure. Omniglot's figures appear to be less conservative and would probably be more justifiable on the assumption that these figures represent the number of people who can speak with native or near native ability. Unfortunately, estimating numbers of speakers of languages is difficult. While official census figures provide useful estimates, not all countries produce relevant figures. In cases where census data includes languages spoken, these estimates can vary widely depending upon the way the census questions are asked and sometimes are affected by a variety of motives on the part of the census takers and the subjects of the census. In the case of minority languages, with a high proportion of bilinguals, many individuals may have high degrees of fluency in

the minority language and may use it in a small range of activities, but in practice may operate almost entirely in the majority language in most situations. This begs the question of whether they are counted in the minority language column or in the majority language. Do we include only native speakers (assuming we could agree on a definition), or do we also include those who have high degrees of second language ability?

Thus, three organisations give three very widely varying estimates. For the calculations hereinafter, the 'mean' population figures, derived from averaging the figures given by Wikipedia, Ethnologue and Omniglot, as shown in the fifth column of the above table will be used for the remainder of the study.

IV. Theoretical Models

While the EDA methodology explicitly avoids preliminary hypotheses, it is useful to have a system by which the data can be presented in an organised fashion. Merely presenting data for each language has no value unless the data is related to something. The point of comparison will be twofold: minority languages will be compared to the majority languages with which they compete and languages will be grouped by status.

Two methods of classifying languages will be used. The first method - 'language constellations' - will group languages into geographic groups reflecting the real-world geographical locations of languages. For example, the languages spoken in the United Kingdom, will be studied in a group that includes all the other languages normally spoken or used in the United Kingdom. The second method will group languages by a newly developed classification system, using more traditional means that are usually employed in minority language studies, where languages are grouped by various objective criteria relating primarily to the status and vitality of those languages. Languages that are the official languages of sovereign states will be examined together, whereas languages that are minority languages will be examined separately.

A. Language Constellations

Languages do not exist in isolation. All languages share geographical and intellectual space with other languages. In many respects languages are in competition with each other. The 'language constellation' system proposed hereinafter is based on an existing model of language classification proposed by Abram de Swaan in 2001(de Swaan, 2001).

De Swann, borrowing heavily from parallels and direct comparisons to economic theories, defines a language as a type of economic 'good', more specifically a 'hypercollective good'. By this he means that human beings make the same choices toward language use as they do with other goods. Following the logic that human beings make a deliberate choice to either continue to use or to acquire a particular language from a set of languages within the geographic area where they live and operate. These languages fall into a

three (or perhaps four) tier hierarchy of languages: 'peripheral', 'central', 'supercentral' (and in the case of increasing dominant English – 'hypercentral'[25]).

The 'peripheral' languages are those that are primarily transmitted within communities, but which have little or no formal recognition on an official level and which are rarely used for communication between groups of persons who do not speak those languages. They are the languages of ethnicity and community and are often employed exclusively by persons whose interpersonal relations do not need to extend beyond that specific community. These would correspond to the minority languages as discussed above.

Above the 'peripheral' languages are the 'central' languages, which are often the national languages of given states. These are the languages that are usually taught in school regardless of the mother tongue of the students. These languages are often given formal recognition by states, oftentimes being the 'official', or *de facto* official, languages of states. Languages such as Danish, Greek and Italian are central languages in their states of origin and are usually learned by all citizens of those states regardless of their mother tongue.

Above these are the 'supercentral' languages, which are themselves central languages, but that are further used to facilitate communication between persons who speak different central or peripheral languages. Supercentral languages are those central languages that have achieved a level of recognition beyond the borders of the state where they are primarily used, and are the principal languages that are learned by second language learners in order to allow such persons to interact outside of their normal states of residence. French or German, while they are certainly central languages in France and Germany (and other countries), are also languages that are sometimes used by persons who speak other central languages. A Pole and a Romanian might prefer to use German or French in their conversation rather than attempt to learn each other's central language. Supercentral languages tend to be local. Historically, German was the supercentral language of central Europe, French the supercentral of western Europe, and Russian the supercentral language of eastern Europe. In other parts of the world different languages tend to become supercentral.

---

[25] De Swaan has not made his mind up on whether English has become the 'hypercentral' language, though it is clear that if English continues to grow, that it will have to be accepted as such. Thus, there are either three or four language classifications in de Swaan's model, depending on whether English is a hypercentral language or not.

Increasingly, however, the supercentral language of choice, in Europe and in the rest of the world, is English. For this reason it is possible to consider English as a special type of supercentral language – a hypercentral language. Increasingly English is replacing many other supercentral languages, as the supercentral language *par excellence*, at least in Europe and increasingly in other parts of the world.

In general, de Swaan believes that these three/four types of languages compete with each other in defined geographic areas. A person inhabiting these areas will make a choice as to whether he continues to speak a particular language, or acquire another language, based on how valuable that language is within the particular area. He calls these language environments 'constellations'.

1. Mapping Constellation S'

De Swaan specifically likens languages to heavenly bodies with the peripheral languages forming the moons, the central languages equivalent to the planets and the supercentral languages the suns. In an effort to make this theory clearer, the following diagrams have been created to this theory.

Figure 32 below represents a situation where four 'peripheral' languages are spoken in a given constellation (S'), say a particular state where several languages (or distinct dialects) are spoken but where the official language of that state is one particular language (or dialect). In some cases, Languages R1 to R4 may be mutually unintelligible, or may be different dialects, though related to C1.

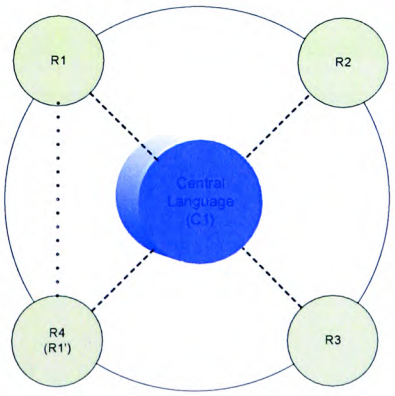


Figure 32 - Peripheral and Central Languages within Constellation S'

This model, while theoretical, reflects broadly the situation in many states today. For example, Spain, France, the UK and Italy all have language situations similar to this model. In Spain, the Central Language, Castilian (Spanish) is used by all organs of the state and is taught to all school children (with certain, recent exceptions in Catalonia). There are four officially recognised "regional" languages (Galician, Basque, Catalan and Valencian), three of which are related to the central language, but generally are not mutually understandable by speakers of the central language only. In the UK, the situation is similar, although there are three spoken minority languages (Welsh, Gaelic and Irish), and two 'revived'[26] regional languages (Cornish and Manx)[27]. However, none of the minority or revived languages in the UK is related to the central languages and is mutually unintelligible, and thus all communication not solely in one peripheral language usually occurs in the central language.

The situation as detailed above is quite common. Prior to the advent of the globalised economy, and prior to the creation of the web, Europe's larger states had succeeded in imposing one central language within each state, while relegating other languages and dialects to peripheral status. A few smaller nations have done likewise, though in some notable exceptions, bilingual or trilingual states have been formed.

2. Mapping Constellation S''

a) S'' with No Supercentral Language

The situation in Figure 32 above is meant to be illustrative of a typical situation within a given state. Once we look at communication within a larger geographic area, the situation becomes a little more complicated. Figure 33 below presents a simplified version of a situation that typically prevails over a larger geographic area, S'', which would represent a larger unit than a state, such as a part of a continent or another equally large area, which area contains a number of states, each one forming its own language constellation (S') as discussed above. S'' is a collection of S' constellations.

---

[27] There are also two varieties of English, Scots and Ulster Scots, that are recognised by the UK government as being regional 'languages' under the terms of the European Charter of Regional and Minority Languages.

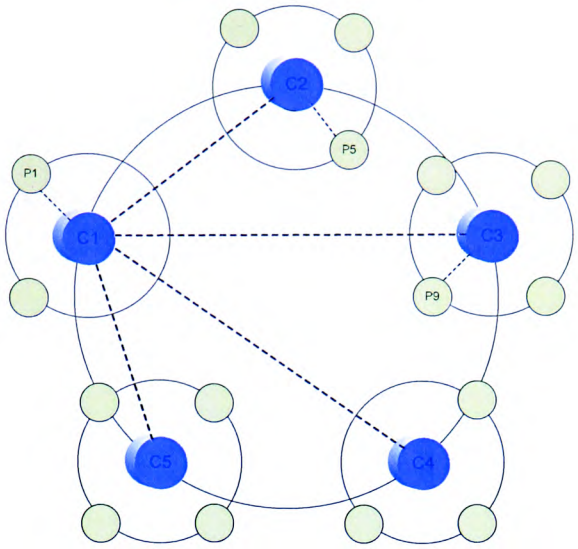


Figure 33 - Language situation without a Supercentral Language within S''

Figure 33 is meant to present a simplified diagram of the situation in western Europe until very recent times. Five S' constellations are shown: the UK, Germany, France Italy and Spain, with English, German, French, Italian and Spanish being the central languages of their S' constellations, and the various regional languages of those states forming the peripheral languages. Obviously a complete mapping of Western Europe would add the constellations of the small nations (e.g. the Netherlands and Portugal).

Any communication between the various S' constellations within S'' would require each party to learn the central language of the interlocutor of the desired constellation (or vice versa). Thus, any speaker of C1 would have to learn 4 other languages (C2 – C5) to be fully conversant with all speakers of the central languages of S''. For a native language speaker of a peripheral language, the onus would be increased by 1 to a total of 5 languages. While not commonplace, it is quite possible to become multilingual to that level. This is mitigated by the fact that full communication within constellation S'' is only necessary for a small number of people, and that oftentimes full communication can be achieved between multilingual speakers who have at least one of the central languages, though not necessarily their native language. Full communication could be had between a subset of multilingual speakers, each with abilities in at least three central languages. While high degrees of multilingualism are not uncommon, it is obvious that the situation represented in Figure 33 is difficult.

b) S'' with a Supercentral Language

Figure 34 shows a situation where there are five central languages within a larger geographic constellation (S''), but where one of the languages has moved into a 'supercentral' position. The situation mapped in Figure 34, for the same language constellation (S''), is therefore considerably simpler for most speakers within the constellation:

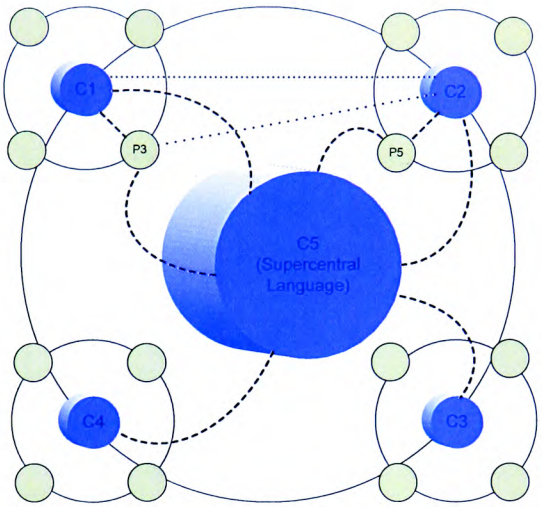


Figure 34 - Central and Supercentral Languages within Constellation S''

Figure 34 could easily represent Western Europe in the present day, with C1 to C4 representing German, French, Italian and Spanish respectively, with C5 representing English.[28] While the language designated as the supercentral language is shown as different, a supercentral Language itself may be a central language within its own S' constellation, and would therefore have its own peripheral languages in 'orbit' around them. Note that the position of languages C1 to C4 is identical to the peripheral languages R1 to R4 in Figure 32. Unless a speaker of C1 to C4 acquires one of the other central languages (represented on the diagram by the dotted line connecting C1 and C2), such a speaker needs only to acquire the

[28] Obviously, this representation is incomplete since the Portuguese, Dutch and Scandinavian situations are not represented, nor are the rather more complex Swiss and Belgian situations. Also, the representation of English (C5) has been simplified in that English's peripheral languages are not shown.

supercentral language to obtain the most efficient communication abilities within this constellation. This is a situation identical to that presented in Figure 32.

By having a language occupy a central position within this constellation (the 'supercentral' position), any mother-tongue speaker of one of the central languages now needs only to learn the supercentral language. Likewise the burden of a peripheral language learner is reduced to learning only the relevant central and supercentral languages (C1 to C4 *and* C5). With a maximum of only 2 additional languages, even a peripheral language speaker can obtain full communication abilities with everyone in the constellation, provided that everyone also learns the supercentral language. Furthermore, with the concentration of resources that arises when a constellation has only one supercentral language, the number of speakers and the availability of materials and learning opportunities increases, making learning and acquiring knowledge of the supercentral language easier.

By enlarging the constellation from S' to S'', pressure is exerted on both the central and more importantly the peripheral languages. The pressure is to reduce the situation to one similar to that which pertains in S' whereby there is a single central language around which a number of peripheral languages orbit.

The language constellation system as outlined above will be used to group the results of the data obtained during the course of this study. This is proposed as one possible method of organising the data for this study and possibly for future studies. Since this study is particularly attuned towards minority languages, it is by comparing minority languages with the very majority languages that they compete with that will show how those languages are doing.

B. Language Classification

There have been a several attempts to classify languages in terms of their vitality, usually defined as whether they are increasing, remaining stable or decreasing in terms of numbers of speakers. The first and still most commonly used is the Graded Intergenerational Disruption Scale (GIDS) proposed by Fishman (Fishman, 1991:87ff). This scale is useful for many purposes, and is based on an assessment of how a language is passed from one generation to another as a spoken language. Fishman designed the GIDS to provide a common framework for measuring minority language vitality and for allowing cross comparisons to be made to other languages. However useful the GIDS is in socio-linguistic studies, it is difficult to apply to the analysis of language use on the web, since we are not studying speakers *per* se, but rather text and language communication divorced from speakers. The GIDS was primarily designed to classify languages along a wide range of real world uses, and as such it is a valuable tool.

	GIDS (adapted from Fishman 1991)
LEVEL	DESCRIPTION
1	The language is used in education, work, mass media, government at the nationwide level
2	The language is used for local and regional mass media and governmental services
3	The language is used for local and regional work by both insiders and outsiders
4	Literacy in the language is transmitted through education
5	The language is used orally by all generations and is effectively used in written form throughout the community
6	The language is used orally by all generations and is being learned by children as their first language
7	The child-bearing generation knows the language well enough to use it with their elders but is not transmitting it to their children
8	The only remaining speakers of the language are members of the grandparent generation

Table 24 - GIDS (Fishman, 1991)

To get over the limits of the GIDS, several other schemes have been proposed. For example, a UNESCO panel proposed a new scheme in 2003 (Brenzinger et al., 2003):

Degree of endangerment	Intergenerational Language Transmission
Safe	The language is spoken by all generations; intergenerational transmission is uninterrupted
Vulnerable	Most children speak the language, but it may be restricted to certain domains (e.g., home)
Definitely endangered	Children no longer learn the language as mother tongue in the home
Severely endangered	The language is spoken by grandparents and older generations; while the parent generation may understand it, they do not speak it to children or among themselves
Critically endangered	The youngest speakers are grandparents and older, and they speak the language partially and infrequently
Extinct	There are no speakers left

Table 25 - UNESCO Classification Scheme 2003

Ethnologue *(Ethnologue.com, 2008)* categorises language vitality in terms of a five level scale. The scale focuses primarily on the number of native speakers than on other factors:

Category	Description
Living	Significant population of first-language speakers
Second Language Only	Used as second-language only. No first-language users, but may include emerging users
Nearly Extinct	Fewer than 50 speakers or a very small and decreasing fraction of an ethnic population
Dormant	No known remaining speakers, but a population links its ethnic identity to the language
Extinct	No remaining speakers and no population links its ethnic identity to the language

Table 26 - Ethnologue Classification Scheme 2008

Lewis has proposed a more expanded system, believing that, as Ethnologue is a widely-used reference volume, it would be advantageous for it to report data using a framework that

represents current best practice and that can be applied consistently to all of the world's languages whatever their degree of endangerment or development (Lewis & Simons, 2009):

Expanded Graded Intergenerational Disruption Scale (adapted from Fishman 1991)			
LEVEL	**LABEL**	**DESCRIPTION**	**UNESCO**
0	International	The language is used internationally for a broad range of functions.	Safe
1	National	The language is used in education, work, mass media, and government at the nationwide level.	Safe
2	Regional	The language is used for local and regional mass media and governmental services.	Safe
3	Trade	The language is used for local and regional work by both insiders and outsiders.	Safe
4	Educational	Literacy in the language is being transmitted through a system of public education.	Safe
5	Written	The language is used orally by all generations and is effectively used in written form in parts of the community.	Safe
6a	Vigorous	The language is used orally by all generations and is being learned by children as their first language.	Safe
6b	Threatened	The language is used orally by all generations but only some of the child-bearing generation are transmitting it to their children.	Vulnerable
7	Shifting	The child-bearing generation knows the language well enough to use it among themselves but none are transmitting it to their children	Definitely Endangered
8a	Moribund	The only remaining active speakers of the language are members of the grandparent generation.	Severely Endangered
8b	Nearly Extinct	The only remaining speakers of the language are members of the grandparent generation or older who have little opportunity to use the language.	Critically Endangered
9	Dormant	The language serves as a reminder of heritage identity for an ethnic community. No one has more than symbolic proficiency.	Extinct
10	Extinct	No one retains a sense of ethnic identity associated with the language, even for symbolic purposes.	Extinct

Table 27 - EGIDS Classification (Lewis 2009)

Pimienta, Prado and Blanco have proposed a classification regime more attuned to the study of languages on the internet (Pimienta et al., 2009):

CATEGORY	ROLE OF THE INTERNET
Main spoken languages	The Internet could play a role of amplifier of presence, especially in a transition period when the repartition of internet users by language is not even due to the digital divide. Note: our thesis is that this transitory period has been over for English few years ago.
Official languages covering more than one developed country (like Italian or Dutch)	There is an opportunity to be seized in the virtual world. The "international" status of those languages shall give trust to the speakers to an easy relation across borders.
Official languages spoken in only one developed country (like Norwegian, Greek, Danish or Japanese)	There is a need for a vigorous virtual linguistic policy to support a presence in the virtual world comparable or stronger than in the real world with long term positive feed-back for the place of the language in the word; speakers however can feel a barrier for international relation using their language.
Local languages of developed countries (like Sardinian, Galician, Welsh, or Frisian)	They are threatened by a strong pressure from both English and their respective national language. The diagnostic is uncertain without virtual linguistic policy and depends on specificities although the case of Catalan is to be followed as a success story, both at virtual and non virtual level.
Lingua franca of speakers of some developing countries (like Hausa, Quechua, Pulaar or Swahili)	A positive future shall be possible whereas the digital divide is really overcome and virtual linguistic policies are defined.
Languages of developing country covering more than one country but only used by native speakers (like Aymara, Guarani or creoles)	Theoretically a positive future should be possible whereas the digital divide is really overcome; however there is a present correlation between lack of access and belonging to indigenous communities which does not give yet signal of changes. The case of Paraguay where Guarani is given instruments following its status of official language is to be followed with interest.
Official languages of a unique developing country (like Slovenian or Albanese)	They are under strong pressure from both English and respective powerful regional languages which could trigger negative prospects in the absence of virtual policy.
Local languages of developing countries (like Chabacano, Maya or Mapuche)	If the language is provided with the appropriate linguistic tools (and first a normalized and stable system for writing and grammar), a , linguistic policy focusing the production of local content could help. However there are not many example today of this kind.
Languages threatened or disappearing (like Ainu)	The Internet could, at worst, become a formidable tool to for conservation of the written or oral patrimony, at best, accelerator of policies for language adaptation.
Languages very seriously threatened or disappearing	The Internet could at least allow to disseminate the patrimony of that language will leave if urgent digitalization campaigns are done..

Table 28 - Classification of Languages (Pimienta et al. 2009)

C. Web Presence Scale Classification System

For the purposes of classifying languages, the classification system proposed by Pimienta, et al. is slightly reworked and a numbering system is added. This reworked scheme is designed to work in conjunction with the web presence scores (as in the next section below), whereby languages are classified into one of ten tiers, according to how much presence they have on the web, using a 'Web Presence scale'. This scale is logarithmic, similar to the Richter scale, which allows the measurement of the presence of any language on a scale of 0 to 1.0, where English, the pre-eminent language of the web is used as the base.

Table 28 overleaf details a proposed classification system along with a set of criteria for classifying languages along with a range of expected scores within the framework of the Web Presence Scale.

For Tier I, we can confidently state that only English can satisfy the criteria. In the future other languages (possible Chinese and Spanish) may equal English, but for the foreseeable future it is believed than English has achieved a special status of being *the* global language, especially if we confine ourselves to its position on the web.

Tiers II, III and IV are the majority languages. What separates the three tiers is whether they are spoken in one state or many and whether they are *linguae francae*, or to use the Language Constellation theory elaborated above, Supercentral or Central languages. A *lingua franca* is essentially a supercentral language as proposed by de Swaan above. Those that are spoken in more than one state, it is believed, should have a stronger presence. That being said, languages such as Japanese are powerful enough that they may have more significant presences than more widely spoken languages.

Tier V and Vl languages are the primary focus of this study. Specifically, the question is posed whether these languages suffer any specific disadvantages because of their lesser status.

Tier	Criteria	Expected Score
I	The pre-eminent language of the Web and the international *lingua franca*	1
II	Official language of a very large state or several large states and used as a regional *lingua franca*	.9 – .999
III	Official language of at least one large sovereign state or several sovereign states, but not a *lingua franca*	.8 - .899
IV	Official language of a sovereign state, with little use outside that state	.7 - .799
V	Language with official status within a region of a sovereign state	.6 - .699
VI	Language with no official status on any governmental level	.5 - .599
VII	Lingua franca of some developing countries	.4 - .499
VIII	Languages of a developing country covering more than one country but only used by native speakers	.3 - .399
IX	Languages of a developing country	.2 - .299
X	Languages that are not used on the Web	<.2

Table 29 - Classification of Languages by Web Presence

as minority languages within their constellations. An interesting question to pose is whether a language's status has a noticeable effect on its web presence. For example, the number of speakers of Icelandic (Tier IV) and Welsh (Tier V) are roughly comparable (300,000 and 500,000 respectively). Would a study of their web presence show that there is noticeable difference, and can this be put down to their status as Tier IV and Tier V languages?

Tier VII languages and especially Tier VIII and below are for languages that, for most purposes, are invisible on the web. There a number of problems studying many of these languages since they often have not yet developed the necessary written conventions(Paolillo & Das, 2006). There are a number of sites that have been designed to catalogue these languages, but they are generally designed for linguistic purposes rather than general use. These languages will not be analysed or discussed in this study.

In Table 30 below, the previous table has been populated with example languages, with the languages that are the target of this particular study indicated in bold characters. As can be seen, languages from only the top six tiers have been chosen, since the target languages of this study are minority languages and those minority languages that fall into tiers V and VI. This list is not meant to be exhaustive. The allocation of any one language to a particular tier is open to debate; this list is simply meant to be a provisional list designed to be tested during the course of this study

Tier	Criteria	Example Languages	Expected Score
I	The pre-eminent language of the Web and the international *lingua franca*	**English**	1
II	Official language of a very large state or several large states and used as a regional *lingua franca*	Chinese, **Spanish**, **French**, Arabic, Hindi, Russian, **German**, Portuguese,	.9 – .999
III	Official language of at least one large sovereign state or several sovereign states, but not operating as a *lingua franca*	Japanese, Indonesian, Urdu, **Italian**, **Dutch**,	.8 - .899
IV	Official language of a sovereign state, with little use outside that state	**Danish**, Swedish, Polish, Slovakian, Thai, Turkish, **Icelandic**, etc	.7 - .799
V	Language with official status within a region of a sovereign state	**Welsh**, **Irish**, **Scots Gaelic**, **Manx**, **Catalan**, **Frisian**, **Sardinian, Galician,** etc.	.6 - .699
VI	Language with no official status on any governmental level	**Breton**, **Cornish**, Occitan, Romany	.5 - .599
VII	Lingua franca of some developing countries	Hausa, Swahili	.4 - .499
VIII	Languages of a developing country covering more than one country but only used by native speakers	Aymara, Guarani or creoles	.3 - .399
IX	Languages of a developing country	Various languages of Africa, Asia, South America and Oceania	.2 - .299
X	Languages that are not used on the Web	The majority of the world's languages, spoken sometimes by very few speakers or which have not yet developed a written form	<.2

Table 30 - Classification of Languages by Web Presence with Example Languages

D. Calculating Web Presence Scores

1. Rejecting Number of Bytes as a Useful Measure

It was mentioned above that depth of content for a Wikipedia edition could possibly be determined by looking at the raw number of bytes that a set of articles of a Wikipedia edition uses. Theoretically, this method should be sound since all information contained in a Wikipedia article would need to be encoded and this would be reflected in the number of bytes used by that article. In order to test this particular measure, a set of samples were obtained on 27 June 2010. The raw calculations for the English language edition, as produced by SPSS, are:

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	3000	3000	3000	3000	3000	3000
	Missing	0	0	0	0	0	0
Mean		53655.97	1468.95	1038.37	3.63	186.99	11.00
Median		35408.50	719.50	319.00	1.00	117.00	5.00
Mode		26572	377	0	0	68	2
Std. Deviation		57100.778	2184.072	2087.950	14.871	212.623	37.031
Variance		3.260E9	4770170.741	4359536.158	221.145	45208.689	1371.259
Range		1204569	21723	21727	528	3978	1762
Minimum		19021	4	0	0	0	0
Maximum		1223590	21727	21727	528	3978	1762
Sum		160967909	4406862	3115095	10883	560960	32985

Table 31 – Statistics for EN

While the Welsh language editions are:

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	3000	3000	3000	3000	3000	3000
	Missing	0	0	0	0	0	0
Mean		25481.62	410.91	185.00	1.17	111.04	2.57
Median		22064.00	314.00	93.00	1.00	89.00	2.00
Mode		16954[a]	278	0	0	47	2
Std. Deviation		18837.656	418.580	398.065	7.097	99.636	1.928
Variance		3.549E8	175209.305	158455.464	50.372	9927.277	3.718
Range		669648	10702	10813	289	3493	27
Minimum		12943	111	0	0	30	2
Maximum		682591	10813	10813	289	3523	29
Sum		76444867	1232738	555009	3504	333124	7723

a. Multiple modes exist. The smallest value is shown

Table 32 – Statistics for CY

The means can be compared as is shown in the following chart (in order to render the chart more readable, the scales of some means have been modified)

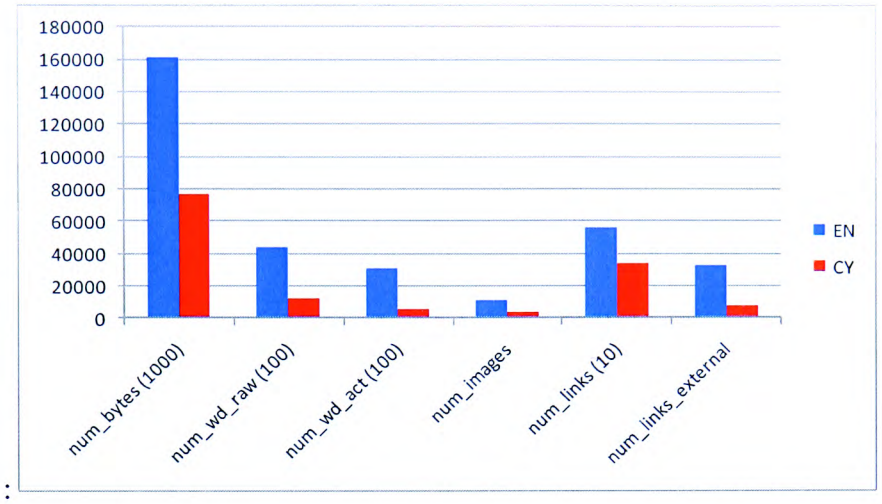


Table 33 – Comparison of English and Welsh Samples

If we look at the number of bytes, it can be readily seen that the mean for the Welsh language edition is approximately half the value for the English language edition. But, when we look at the values for numbers of words, a more marked divergence between the two means can be noted. In general, it is clear that a Wikipedia article contains extraneous data that shows up in the byte count, but does not necessarily translate into informational content within the article. For this reason, it would seem best to disregard the number of bytes per article as a valid basis of comparison, and instead concentrate on the other factors that present a more nuanced and possibly more accurate measure of the contents of a Wikipedia article. The more important points of comparison are the number of words, the number of images and the number of links.

2. Calculating WikiScores

As stated above, languages will be classified in two different manners: language constellations and a tiered language classification scheme. In order to provide the numbers for comparison a basic score will be calculated for each language based on that language's presence on Wikipedia. Two basic formulae will be proposed for the calculation of these scorse: a 'Raw Wikiscore', which is essentially the sum of the various components of an

average article multiplied by the number of articles in the Wikipedia database. The second will be a 'Population Wikiscore', which will take the raw Wikiscore and divide it by the number of speakers of a particular language, in an attempt to analyse the various language editions on an equal basis, as has been done in a number of studies and in Wikipedia's own statistics.

3. Raw Wikiscore

Raw Wikiscores are intended to measure the actual information contained in a Wikipedia. The score is a composite of various components of an article, being the three important content items of a Wikipedia article: the article's actual verbal content (num_words_act), the content images (num_images) and the number of links provided in the article to other links on the web (num_links). The number of words, images and links are then weighted, as per the following basic formula:

$$WS_{raw1} = N\left(\left(A \cdot w_{words}\right) + \left(I \cdot w_{images}\right) + \left(L \cdot w_{links}\right)\right)$$

Equation 1 - Raw Wikiscore

Where N is the total number of the articles in a particular language edition of Wikipedia, A is the average number of actual worlds multiplied by the word weighting for that language (w_{words}),) I is the average number of images multiplied by the weighting assigned to images (w_{images}), and L is the average number of external links multiplied by the weighting assigned to links (w_{links}).

A second version of the formula is also proposed, where the number of external links is used, so as to better reflect the greater usefulness of an article that is better referenced by providing the user more information, as per the following formula:

$$WS_{raw2} = \frac{N\left(\left(A \cdot w_{words}\right) + \left(I \cdot w_{images}\right) + \left(L_e \cdot w_{extLinks}\right)\right)}{100000}$$

Equation 2 - Raw Wikiscore

Since the results obtained by this formula give extremely large results, a simple factoring by 100,000 gives a more readable result, as per the following formula:

$$WS_{raw1} = \frac{N\left(\left(A \cdot w_{words}\right) + \left(I \cdot w_{images}\right) + \left(L \cdot w_{links}\right)\right)}{100000}$$

Equation 3 - Raw Wikiscore (factored)

The number of articles for a particular language edition of Wikipedia is given on that language edition's homepage in precise numbers, and is updated daily. However, Wikipedia provides an historical list of monthly article counts[29], with the actual counts rounded off. Since it is a time consuming process to go to each language edition's homepage, and given that such numbers change from day to day, it was decided to use the less precise but more accessible numbers provided by Wikipedia's historical summary. This has the additional advantage that such article counts are easily verifiable. For the various calculations that follow, the following article counts were used, which were the article counts given by Wikipedia's article count page for September 2010:

Language	Article Count	Language	Article Count
EN	3,500,000	BR	36,000
DE	1,200,000	NL	656,000
FR	1,000,000	FY	18,000
IT	758,000	FRR	1,000
ES	676,000	STQ	2,000
CY	30,000	DA	139,000
GA	12,000	IS	30,000
GD	8,100	CA	296,000
GV	3,600	GL	65,000
KW	2,000	SC	2,500

Table 34 - Wikipedia Article Counts

[29] http://stats.wikimedia.org/EN/TablesArticlesTotal.htm

a) Word Weightings

Some languages use more words than others to convey the same amount of information. For this reason it was decided to weight the number of words using an objective criterion. The weighting (w_{words}) is based on the total number of words that a particular language needs to express the same information. To this a sample text was sought that would give the same content, but in all languages. While it is simple to find such a text for widely spoken languages (translations of international legal documents are one good source) it was difficult to find a suitable text that was also translated into the lesser-used languages. Fortunately, Omniglot provides a version of the Bible, Genesis 11: 1-9[30]. In each case, where different Bible versions are provided by Omniglot[31], the lastest version has been chosen. Admittedly, these are not ideal, but they do give an independent and verifiable metric.

Language	Num Words	Weighting
EN	222	1.000
DE	190	0.856
FR	207	0.932
IT	194	0.874
ES	207	0.932
CY	176	0.793
GA	224	1.009
GD	231	1.041
GV	238	1.072
KW	207	0.932
BR	207	0.932
NL	227	1.023
FY	212	0.955
FRR	212	0.955
STQ	212	0.955
DA	212	0.955
IS	199	0.896
CA	201	0.905
GL	181	0.815
SC	194	0.874

Table 35 - Omniglot Word Counts

[30] The index page for the different versions are available at http://www.omniglot.com/babel/index.htm

[31] For example, for English, Omniglot provides seven different translations over a long time period (Wycliffe version (1395), Tyndale version (1536), King James version (1611), Douay Rheims version (1899), Basic English version (1949), New International version (1973) and English Standard Version (2002).

It should be noted, that no versions were given by Omniglot for North Frisian (FRR), Saterland Frisian (STQ) and Sardinian (SC), and the value for West Frisian (FY) was used for these two dialects of Frisian. Likewise as no Sardinian version was available, the value for Italian was used for Sardinian.

This is not entirely a successful exercise, since there are a number of potential problems in using such a small word set and the problem is further compounded in that each translation would was produced by a different translator. Translation is not an exact science, and two different translators could, and often will, come up with different ways of translating any given text. This particular source of obtaining a word weighting is, therefore, not entirely satisfactory. However, as is shown in Table 35 above, there is not a great deal of discrepency between the different language versions, with a range of some fifty words between the smallest and largest versions, and an average of about 207 words. Thus, this method has been retained until a more accurate source can be found.

It should be noted that given the small differences that exist between the various languages, it might be possible to do away with a word weighting altogether. However, all the languages that have been the target of this study are languages of Western Europe, and are from only three language families (Germanic, Celtic and Romance). A great deal of discrepency would be unusual. If this study is extened to languages that are not from these three families, and indeed is extended to languages that are not Indo-European, we might see a greater range of difference, and thus word weighting may be more important.

b) Image weightings

It is difficult to determine accurately what weighting should be assigned to an image contained in a Wikipedia article. On the one hand, it seems obvious that we cannot simply count an image as being equal to one word, since it seems self evident that an image gives more information than a single word could. On this basis, an initial reaction to assign a weighting factor of 1000 would work, on the basis that a "picture is worth a thousand words". However, when the various means of the Wikipedia editions were analysed it appeared that for all language editions of Wikipedia, the mean word count was measured in the hundreds, and in the case of minority languages, in the low hundreds (e.g. English – 948.41; German – 558.2; Welsh – 191.70; Irish – 319.9), and, with a 1000 factor weighting, a single picture would carry three to ten times as much value as the content of the article. Other numbers were

tested, with a decision that a weighting of 50 would be appropriate. It should be noted that, at the time of writing, any weighting tends to increase the Wikiscore of the larger languages, simply because they contain more images. However, it must be noted that a value of 50 for the image weighting is entirely arbitrary and is open to discussion, testing and further analysis.

c) Link Weightings

The same problem exists for the choice of an appropriate weighting for links as for the weighting for images. Since links to other areas of Wikipedia are of lesser value, it was decided, on an arbitrary basis, to use a value of 10 for external_links. External links are oftentimes of great informational importance and tend to increase the usefulness of an article greatly, by providing a range of additional information for the user. Indeed, one could argue that some external links are much more useful than many pictures. However, in general, it is believed that many such external links are of the footnote variety and do not necessarily add much more information. For this reason a compromise was decided upon whereby external links would be weighted at 20. As with the value for image weighting, the choice of a weighting factor is entirely arbitrary and is open to debate, discussion and further testing and refinement.

4. Wikiscore by Population

Given the belief that it is unreasonable to expect a small language with a small user base to produce the same amount of material, the Wikiscores will also be factored by the number of speakers of the particular language. As with the raw Wikiscores two slightly different formulations are given: WS_{pop1} and WS_{pop2}:

$$WS_{pop1} = \frac{WS_{raw1}}{P}$$

Equation 4 - Wikiscore by Population = Version 1

And a second, slightly different version, using the raw2 version of Wikiscore:

$$WS_{pop2} = \frac{WS_{raw2}}{P}$$

Equation 5 - Wikiscore by Population = Version 2

Where P is the estimated population that speaks the language of the Wikipedia edition in question. The figures eventually used for P are those that were used for analysing the article counts by population and are shown in Table 23 - Estimate of World Language Populations on page 80.

Wikiscores are interesting in their own right in that they provide an concrete numerical value that can be used to analyse and chart the progress of any given language edition of Wikipedia, as well as to provide comparisons to other language editions. Wikiscores, if calculated and shown over time, will give an idea of as to the absolute and relative growth of any particular language edition of Wikipedia.

5. Presence Values

The basic formula for determining the presence value for a particular language is a ratio calculated by taking the logarithm of the raw Wikiscore for the target language (YY) and dividing it by the logarithm of the raw Wikiscore of a base language (XX), as per the following formulation:

$$\frac{\log_{10}\left(WS_{raw1(YY)}\right)}{\log_{10}\left(WS_{raw1(XX)}\right)}$$

Equation 6 - Presence Values

In full, the formula is as follows:

$$\log_{10}\left(\frac{\left(\frac{N\left(\left(A\cdot w_{words}\right)+\left(I\cdot w_{images}\right)+\left(L\cdot w_{links}\right)\right)}{1000000}\right)_{(YY)}}{\left(\frac{N\left(\left(A\cdot w_{words}\right)+\left(I\cdot w_{images}\right)+\left(L\cdot w_{links}\right)\right)}{1000000}\right)_{(XX)}}\right)$$

Equation 7 - Presence Values - Full Formula

Two examples can be given: if we calculate the raw Wikiscores for English as 10,536.766, for French as 2,303.166 , and for Welsh as 40.032, we can calculate the presence values for these languages as follows:

Theoretical Models

For French:

$$\frac{\log_{10}(2303.166)}{\log_{10}(10536.766)} = .836$$

And for Welsh:

$$\frac{\log_{10}(40.032)}{\log_{10}(10536.766)} = .398$$

These numbers are on the logarithmic scale and show orders of magnitude rather than direct linear comparison. The score for Welsh is significantly smaller in real terms, than that of French.

These presence scores will be used in the next section to analyse the various languages which we will organise by using the language Constellations.

V. Measuring Language Presence on Wikipedia

Twenty language editions of Wikipedia were chosen as the targets to test the models and calculations outlined in the previous two chapters. The languages were chosen primarily because they are a mix of supercentral, central or minority or regional languages spoken in Europe. The following table lists the languages, along with the two letter language codes as used by Wikipedia:

BR	Brezhoneg (Breton)	GA	Gaeilge (Irish)
CA	Català (Catalan)	GD	Gàidhlig (Scottish Gaelic)
CY	Cymraeg (Welsh)	GL	Galego (Galician)
DA	Dansk (Danish)	GV	Gaelg/Gailck (Manx)
DE	Deutsch (German)	IS	Íslenska (Icelandic)
EN	English	IT	Italiano (Italian)
ES	Español/Castellano (Spanish)	KW	Kernewek (Cornish)
FR	Français (French)	NL	Nederlands (Dutch)
FRR	Nordfriisk/Nordfrasch (North Frisian)	SC	Sardu (Sardinian)
FY	Frysk (Frisian)	STQ	Seelterske/Seetlerfräiske (Saterland Frisian)

Figure 35 - ISO 639-1 language codes - 20 Languages

The data will be analysed using the theoretical model as detailed above. All of the data were obtained from scrapes of 3000 samples for each of the different language editions run during the last two weeks of November and the first two weeks of December 2010, with the exception of Danish and Icelandic, which were carried out in March of 2011. The raw statistical results were produced using SMSS version 14, and the salient data is reproduced in an Appendix.

A. Analysis by Constellations

In this Section, the two principal formulas: Wikiscores and Presence Values will be applied to four language constellations of Western Europe. The following constellations will be given full treatment:

1. S'' – European Union – Major Languages only: EN, DE, FR, IT, ES
2. S' – The United Kingdom and the Republic of Ireland (UKI): EN, CY, GA, GD, GV, SC
3. S' – Germany and the Netherlands (GEN): DE, NL, FY, FRR, STQ
4. S'' – EU – All twenty languages studied

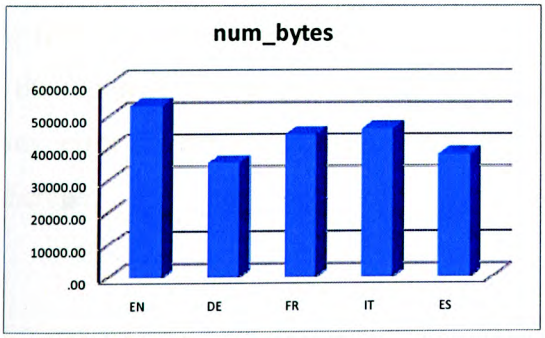
1. Analysis of the Raw Data – Major Languages of the EU

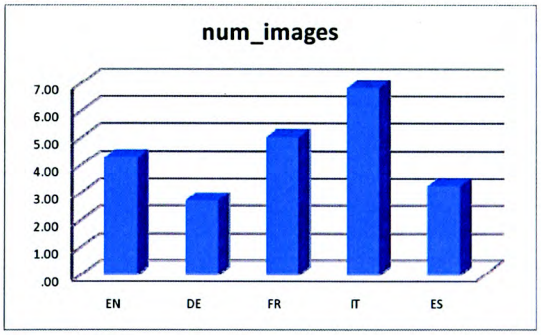
The first test was performed on the Wikipedia editions for five of the major European languages: English (EN), German (DE), French (FR), Italian (IT) and Spanish (ES). The following table shows the means for six of the measurable data types that are the target of this study:

	num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
EN	53029.89	1375.03	948.41	4.27	184.85	10.64
DE	35263.16	820.72	558.29	2.69	107.28	4.83
FR	43925.31	1066.98	628.29	5.01	146.69	5.18
IT	45610.68	1009.75	609.75	6.80	167.53	4.60
ES	37942.39	1047.79	714.57	3.24	120.50	6.16

Table 36 - Means of European Major Languages

Which are graphed as follows:





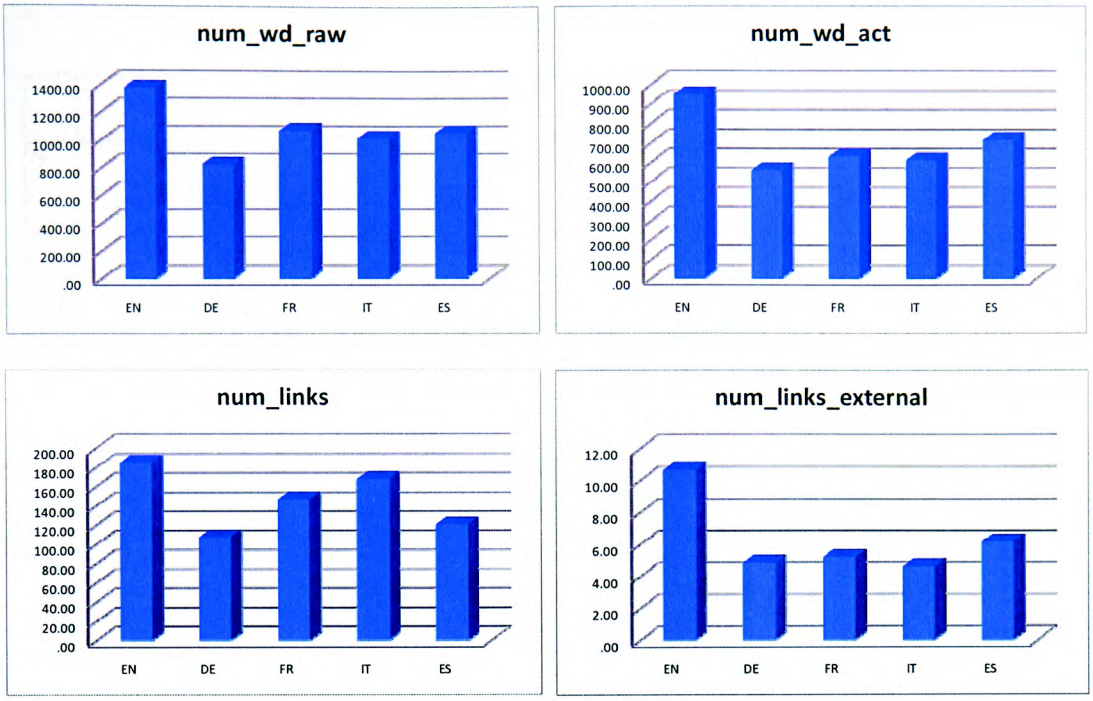


Figure 36 - Means of Samples - Major Languages

These graphs produce some surprising results. The first is that the averages for German are somewhat on the low side, while the results for Italian are sometimes on the high side. This may have been due to an inaccurate sampling, and it order to eliminate this as a potential problem; two additional samples were taken for both languages. The results for the three samples taken from the DE edition are given on the following page:

As can be seen, the results of the sampling are fairly consistent, especially the values computed for the mean, where the number of bytes per page is ~35,500 words per article. The same results were obtained for the three samples of Italian. Thus, there is no reason to believe that the reported differences between the German and Italian editions of Wikipedia are related to any error in the sampling, or as a result of an invalid sample, or any other fault in WkScrape.

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	2935	2935	2935	2935	2935	2935
	Missing	0	0	0	0	0	0
Mean		35263.16	820.72	558.29	2.69	107.28	4.83
Median		28931.00	519.00	268.00	1.00	85.00	3.00
Mode		23615[a]	264	0	1	58[a]	1
Std. Deviation		23908.491	1173.813	1105.976	8.025	85.648	13.227
Range		379533	21524	20890	211	1351	556
Minimum		19067	146	0	0	38	1
Maximum		398600	21670	20890	211	1389	557
Sum		103497362	2408804	1638568	7900	314881	14173

Table 37 - DE Sample 1 (3000 Articles)

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	2925	2925	2925	2925	2925	2925
	Missing	0	0	0	0	0	0
Mean		35688.02	839.58	571.93	2.56	111.21	4.70
Median		29003.00	528.00	272.00	1.00	84.00	3.00
Mode		24172[a]	513	0	0	67	1
Std. Deviation		22895.396	1120.035	1062.728	7.254	92.279	5.214
Range		334581	16403	16544	143	1433	91
Minimum		19094	141	0	0	37	1
Maximum		353675	16544	16544	143	1470	92
Sum		104387462	2455783	1672893	7489	325285	13748

Table 38 - DE Sample 2 (3000 Articles)

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	2930	2930	2930	2930	2930	2930
	Missing	0	0	0	0	0	0
Mean		35899.35	848.31	584.25	2.90	109.26	4.61
Median		28855.00	520.00	271.00	1.00	84.00	3.00
Mode		19976[a]	446	0	1	40	1
Std. Deviation		25276.540	1146.067	1075.075	10.114	92.320	6.152
Range		368881	16885	16963	256	1306	112
Minimum		19213	137	0	0	38	1
Maximum		388094	17022	16963	256	1344	113
Sum		105185098	2485550	1711867	8507	320130	13500

Table 39 - DE Sample 3 (3000 Articles)

The next item of note is to compare results obtained for the number of bytes per article with the other areas measured. While the mean number of bytes per article should be generally speaking a good measure of how much information is contained in each article, it does conceal the true picture. If we compare the number of bytes with images, we can see that while the average Italian article contains roughly the same number of bytes as a French or Spanish article, the average number of images is greater than either and almost double that of a Spanish article.

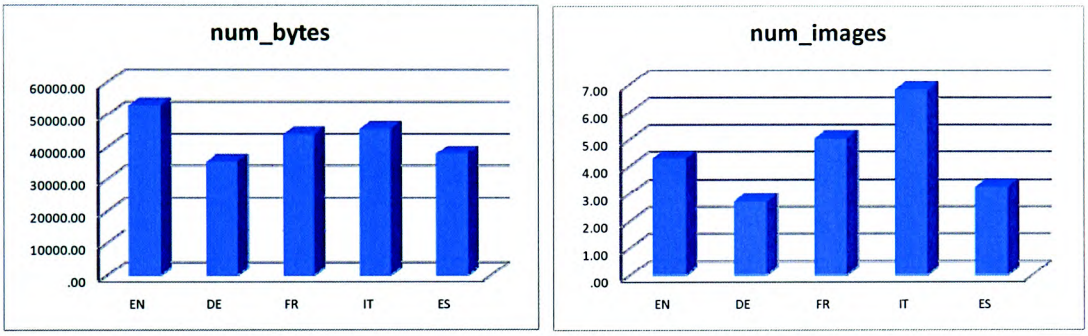


Figure 37 - Num_bytes and Num_images

Likewise, the higher average number of bytes per article of the average English article does not bring out the fact that the average English language article contains nearly twice the number of external links as does an article in the four other languages.

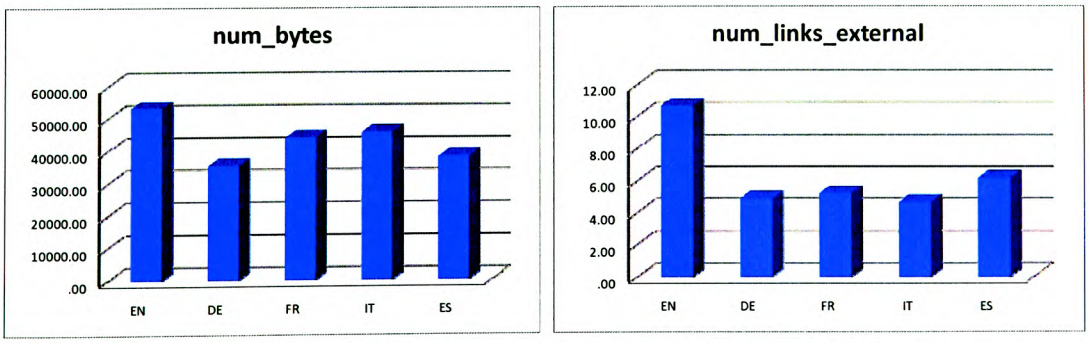


Figure 38 - Num_bytes and Num_links_external

Thus, the average number of bytes per article does not give an entirely accurate picture of what is contained in the average article, as was discussed on page 96 and following.

2. S'' Constellation – The EU – Major Languages

a) Wikiscores

The raw Wikiscores obtained for the five major languages (English (EN), German (DE), French (FR), Italian (IT) and Spanish (ES)) are as follows:

Language	Raw 1	Pop 1	Raw 2	Pop 2
EN	53,608.917	66.021	44,392.297	54.670
DE	13,785.825	104.359	8,507.685	64.403
FR	15,697.135	88.385	9,398.627	52.920
IT	12,967.499	202.934	7,315.987	114.491
ES	9,672.214	23.288	6,431.845	15.486

Table 40 - Raw Wikiscores (S'' EU – Major Languages)

We can attempt to graph these in various ways. The following bar chart gives a good overall picture of the relative sizes of the five different language editions of Wikipedia:

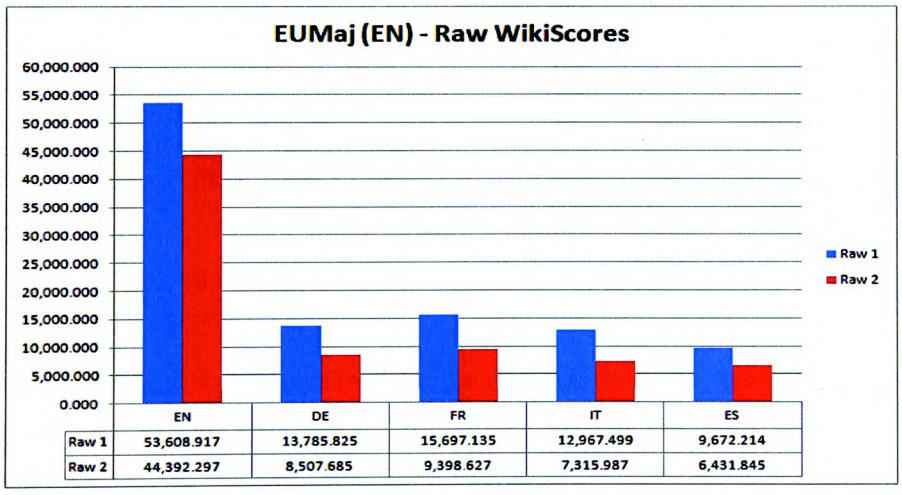


Figure 39- S'' (EU) - Raw Wikiscores

As discussed previously, while there is a significant difference between the WS_{raw1} and the WS_{Raw2} values, they produce, roughly, the same relative results. While the actual numbers change, the relative position of the Wikipedia editions does not.

As can been seen from the chart, it is quite clear that the English language edition of Wikipedia is nearly four times larger than the other major European languages. While the German, French and Italian editions are roughly comparable in terms of size, they are languages mainly spoken by roughly the same numbers of speakers in three of the four largest European states, and we might expect this result. There is a notable degree of difference with respect to the Spanish edition, which is spoken in a sizable European nation and is the principal language of most of South America. While the reasons for this are not the subject of this study, it is clear that Spanish is 'underperforming'.

Other than this remark, the above results are not out of line with what we would ordinarily expect, and this tends to generally confirm the validity of Wikiscapes scraping and analysis algorithms. It is interesting to compare this with the results obtained from Wikipedia's statistics:

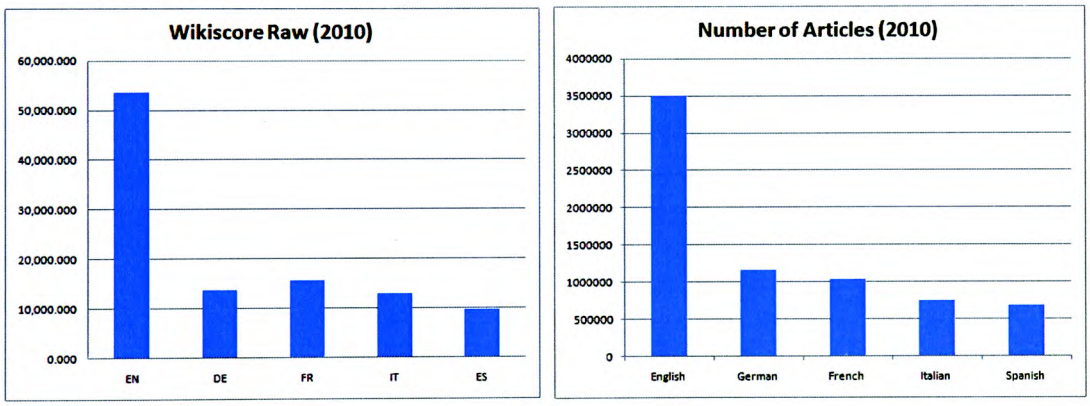


Figure 40 - Comparison of Wikiscores with Number of Articles

The graphs presented in Figure 40 show the results of comparing Wikiscores against the normal statistics provided by Wikipedia, which uses only article counts. On one level, there is a good correspondence in that the relative shape of the graphs is the same. Article counts therefore can provide a good measure of the relative size of a particular language edition of Wikipedia. But as can be seen, there is a significant difference if we look at the results of the German language edition. By article count, German is the second largest Wikipedia edition of those studied. But by using Wikiscores the German edition has less overall content than French and is roughly equal to Italian. This is mainly due to the fact (not revealed in a simple

article count) that the average German article does not contain nearly as much information as the average French article, as can be seen by a study of the raw data in Appendix A.

Even though the difference may be small, it is believed that Wikiscores produce a more accurate picture of the actual contents of a Wikipedia database than does a simple article count. This more accurate picture may be more important if we look at the results of minority languages, since minority language editions of Wikipedia tend to have much less content. This will have the effect of showing the relative sizes of Wikipedia more accurately than will a simple article count.

If we turn our attention to an analysis of Wikiscores by population, we can produce the following bar chart:

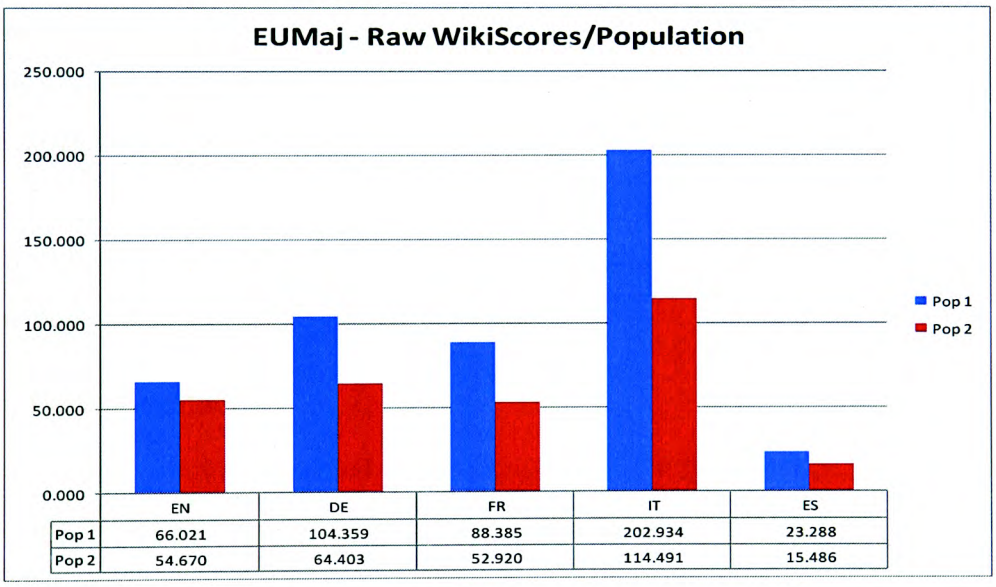


Figure 41 - S" (EU) - Raw Wikiscores/Pop

This produces a rather surprising result, in that the Italian language community is nearly twice as productive per person when compared with the German and French communities and nearly four times more productive per person than the English language community.

Again, an interesting comparison can be made with a similar population comparison using the article count provided by Wikipedia:

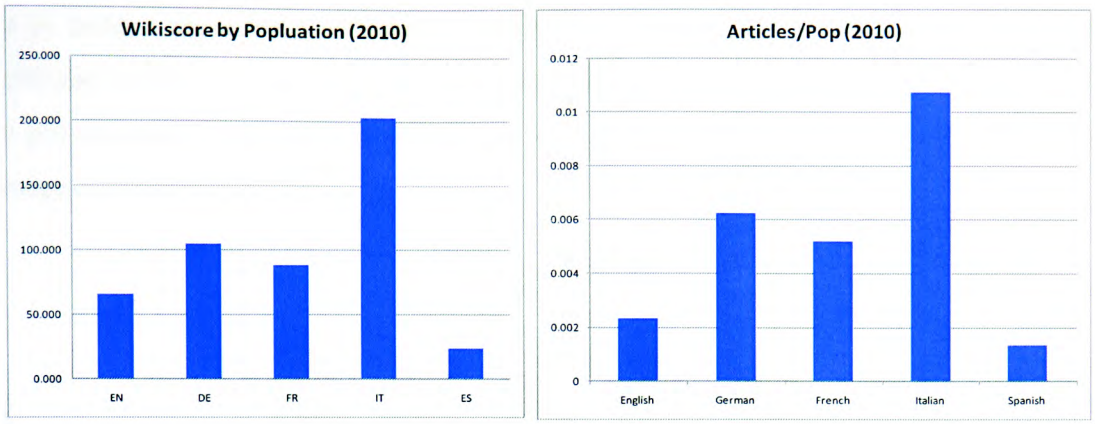


Figure 42 - Comparison of Wikscores/Pop with Articles/Pop

Figure 42 again gives the same overall shape as does the Wikiscore/population graph. However, it can again be noted that there are some differences, mainly in that the Wikiscore/population result for English is not nearly as low as that given by an article count. While this may be a small point, it is believed that given that Wikiscores take into account both the depth of content, that this is a more accurate result.

However, as will be seen below, there is a power law relating to the analysis of the language editions of Wikipedia by population. It is almost impossible to avoid the conclusion that a 'by population' analysis is almost worthless.

b) Presence values

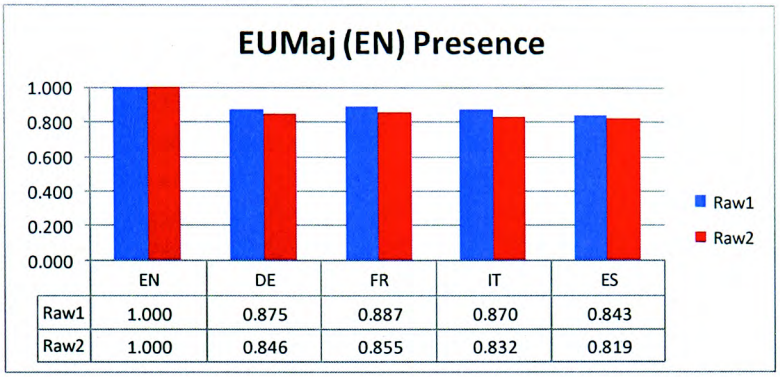
Based on the above calculated Wikiscores, we can now calculate, using the English values as the base, the following presence values for the five language editions of Wikipedia as follows:

Language	Raw1	Raw2
EN	1.000	1.000
DE	0.875	0.846
FR	0.887	0.855
IT	0.870	0.832
ES	0.843	0.819

Table 41 - S'' (EU - Maj) - Presence values

As can be readily seen, if English is used as the base, the other four languages have fairly significant presence values, although we must bear in mind that this is a logarithmic scale and

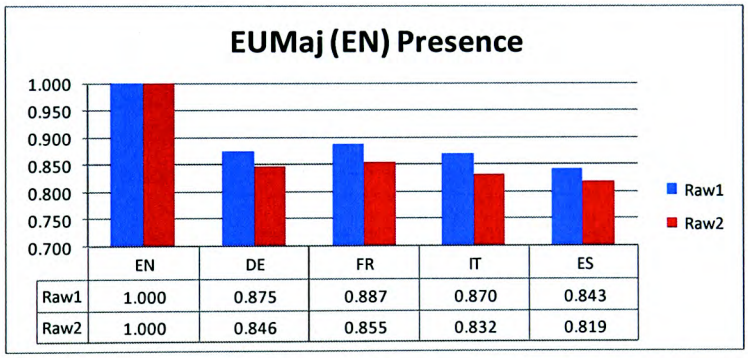
there is quite a significant difference between .887 and 1.000. For this reason, charting logarithmic values can often produce some misleading results. For example, the following bar chart provides an accurate picture, but tends to distort the relative size of the five Wikipedia language editions in question:



	EN	DE	FR	IT	ES
Raw1	1.000	0.875	0.887	0.870	0.843
Raw2	1.000	0.846	0.855	0.832	0.819

Figure 43 – S'' (EU - Maj) - Presence values (version 1)

A slightly better version is given if we only look at the tops of the bars:



	EN	DE	FR	IT	ES
Raw1	1.000	0.875	0.887	0.870	0.843
Raw2	1.000	0.846	0.855	0.832	0.819

Table 42 - S'' (EU - Maj) - Presence values (version 2)

The same information can also be presented in a radar graph

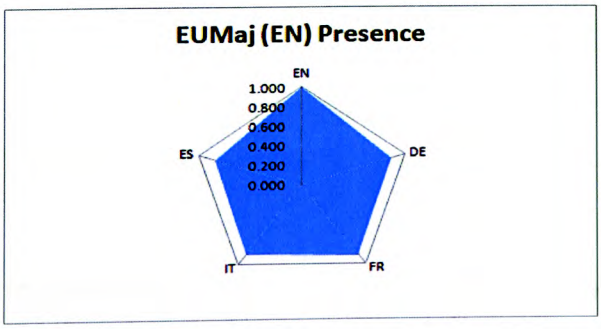


Figure 44 - S'' (EU - Maj) - Presence values (radar – raw1)

But the clearest picture of all is given by a bubble graph:

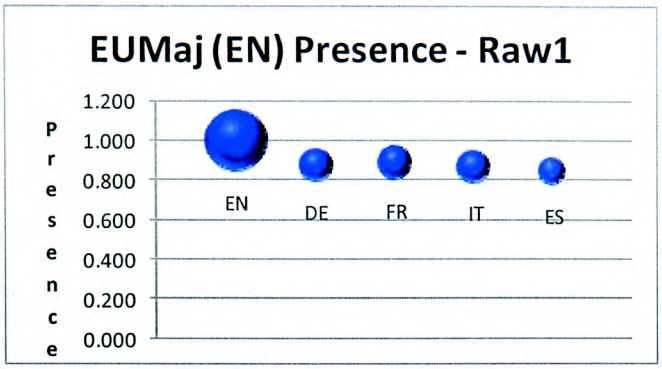


Figure 45 - S'' (EU - Maj) - Presence values (Raw1)

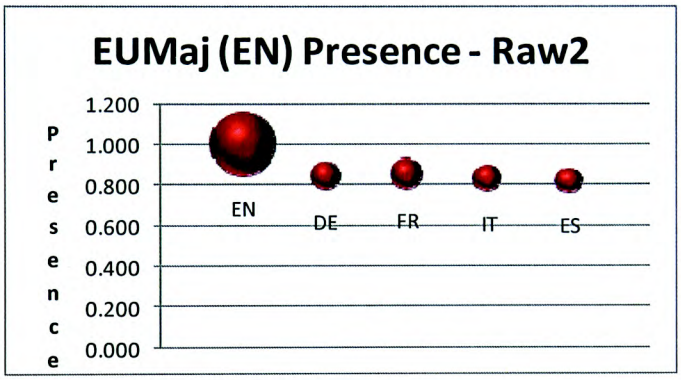


Figure 46 - S'' (EU - Maj) - Presence values (Raw2)

What is shown in the bubble graph is two pieces of information: the y-axis shows the Presence value, while the size of the bubble shows the Wikiscore values – the bigger the bubble the greater the Wikiscore.

The graphs above show us essentially a picture we have already seen in the literature review or can be obtained more easily by a quick analysis of the article counts by language edition already provided by Wikipedia. Additionally, the above graphs do generally demonstrate the reliability of the WkScrape program and the validity of using Wikiscores. By confirming a picture we already know, we can be confident that Wikiscape's algorithms are generally accurate.

But, the real value in developing WkScrape and producing these Wikiscores is in looking at the languages that are not usually the subject of such in-depth analysis: the minority

languages of Western Europe. Given that we have demonstrated and proven the reliability of the results returned, we can now turn our attention to the main focus of this study: minority and regional languages.

3. S' - The United Kingdom and Ireland

For this study, the following languages are contained in the UK and Irelands constellation[32] (S' UKI): English (EN), Welsh (CY), Irish (GA), Scots Gaelic (GD), Manx (GV) and Cornish (KW). These languages have been designated by the UK as minority languages entitled to protection pursuant to the European Charter of Regional and Minority Languages.[33]

a) The Raw Data

The raw data for the means of the various attributes:

	num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
EN	53029.89	1375.03	948.41	4.27	184.85	10.64
CY	28759.90	416.28	191.70	1.53	110.59	2.62
GA	32907.63	590.05	319.90	2.38	122.12	2.82
GD	29225.70	337.81	113.16	1.10	117.52	2.80
KW	34072.58	391.23	56.33	2.10	151.25	2.21
GV	45402.27	561.83	275.29	2.84	194.30	3.49

Table 43 - Statistics for S' UKI Constellation

The conclusions from Table 43 are fairly easy to draw: the Celtic languages in the UKI constellation contain much fewer words, images and links than a corresponding EN language article. The results of this in terms of the Wikiscores will be fairly clear.

---

[32] Some might object to such an artificial constellation, but there are numerous arguments in favour of combining these two sovereign nations into one linguistic constellation: historical, and cultural. There are many points of contact between the cultures. Furthermore, Irish is a language with official status in the Republic and in Northern Ireland. It is also a Celtic language as are all the other minority languages within the UK.

[33] The UK has also designated two variants of English (Scots and Ulster-Scots), but it has been decided to not discuss these two dialects of English.

One point is worth mentioning: the number of bytes is fairly high considering the other numbers. This can be seen clearly if we radar graph the number of bytes with the number of actual words contained in an average article.

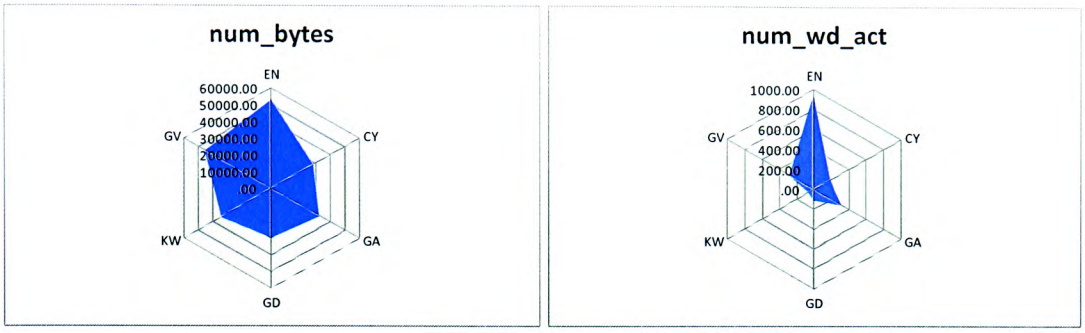


Figure 47 - S' (UKI) Comparison of Bytes/Words

And for comparison, we can look at the same graph produced above for the major European languages:

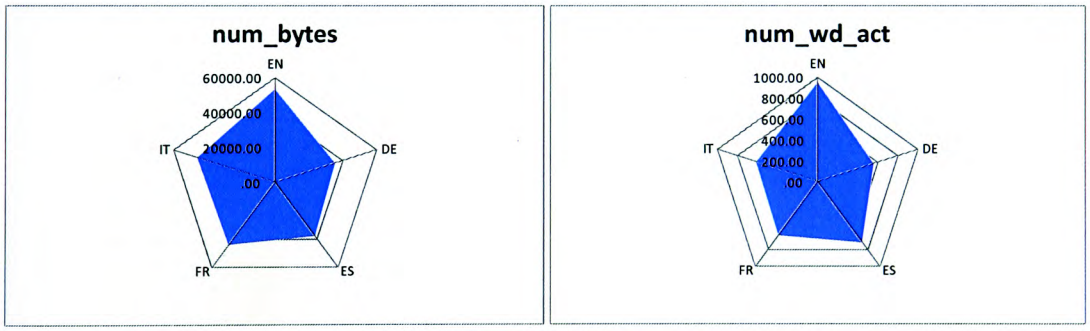


Figure 48 - S'' (EU Maj) Comparison of Bytes/Words

What can be concluded from this is that the number of bytes that any given article uses is fairly constant among the different language editions. Each Wikipedia article contains a lot of non-information code and this tends to be uniform across the language editions. By now it is possible to conclude that the use of the number of bytes of a Wikipedia as an accurate measure of a Wikipedia article can be eliminated from discussion.

b) Wikiscores

As is predictable from the results of the raw data and from knowledge of the number of articles contained in each language edition of Wikipedia, the Wikiscores for the five Celtic languages are not impressive:

Language	Raw 1	Pop 1	Raw 2	Pop 2
EN	53,608.917	66.021	44,392.297	54.670
CY	234.429	337.213	84.251	121.190
GA	126.314	297.332	59.800	140.765
GD	61.599	939.106	18.534	282.560
GV	50.708	1,906.806	18.247	686.163
KW	18.275	4,249.979	4.033	937.995

Table 44 - S' (UKI)- Raw Wikiscores

As can be seen, the scores are really quite low. If we attempt to graph these on a bar chart, we have some difficulties in even showing the data for the five Celtic languages:

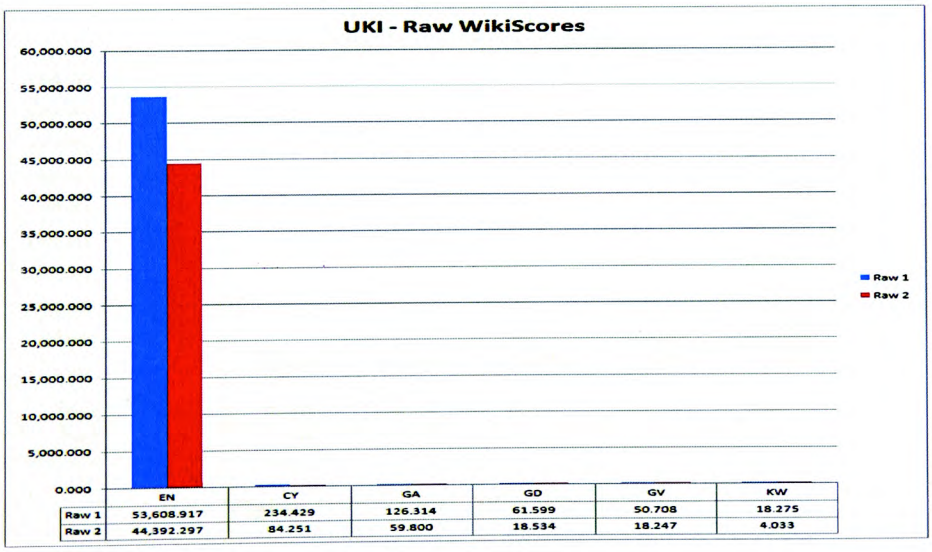
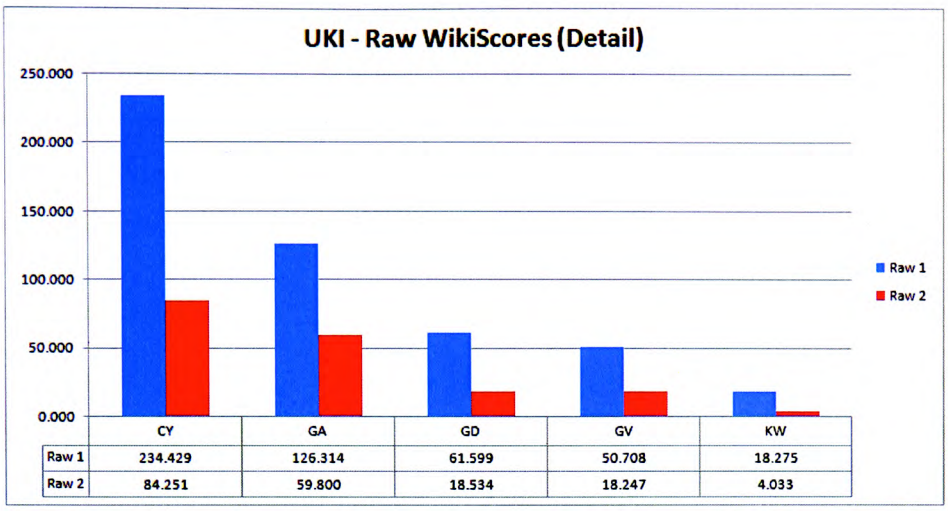


Figure 49 - S' (UKI)- Raw Wikiscores

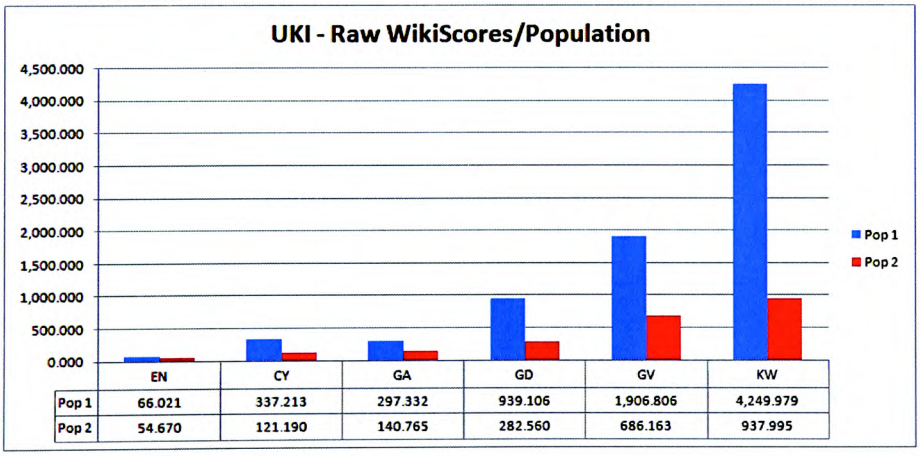
In order to even see and compare the five Celtic languages, we have to remove the English language results to see the details of Celtic language presence:



	CY	GA	GD	GV	KW
Raw 1	234.429	126.314	61.599	50.708	18.275
Raw 2	84.251	59.800	18.534	18.247	4.033

Figure 50 - S' (UKI Celtic Only) - Raw Wikiscores

It is therefore obvious that the Celtic language editions of Wikipedia are very, very small when compared against English. We might, however, draw a different conclusion if we look at the Wikiscores by population:



	EN	CY	GA	GD	GV	KW
Pop 1	66.021	337.213	297.332	939.106	1,906.806	4,249.979
Pop 2	54.670	121.190	140.765	282.560	686.163	937.995

Figure 51 - S' (UKI) - Raw Wikiscores/Pop

However, this is clearly an absurd result. What we have here is a similar, but more magnified, version of the phenomenon seen above when we looked at the EU major languages: the larger the language in terms of population, the smaller the output per person, and vice versa. Simply put, a language with a small population can easily produce some

impressive statistics by virtue of being so small. But the Cornish language edition of Wikipedia only produces a Wikiscore of 18.275 when compared against +53,000 of English. It is hard to argue that the Cornish language edition of Wikipedia is doing well, despite the obvious efforts of a number of dedicated individuals to provide some content in that language. As with number of bytes, it is worth rejecting this metric as an accurate method of measuring a particular language edition of Wikipedia.

c) Presence values

From the calculation of the Wikiscores we can then calculate the presence values of the UKI constellation as follows:

Language	Raw1	Raw2
EN	1.000	1.000
CY	0.501	0.414
GA	0.444	0.382
GD	0.378	0.273
GV	0.361	0.271
KW	0.267	0.130

Table 45 - S' (UKI) - Presence values

Which can be graphed as follows:

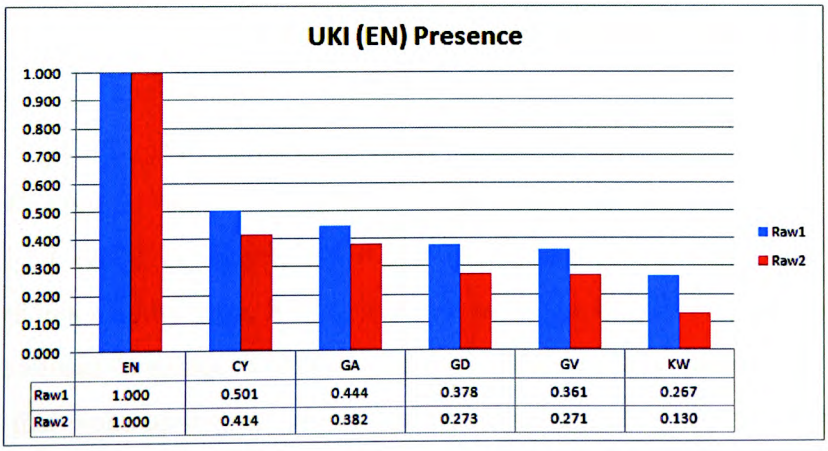


Figure 52 - S' (UKI) - Presence values – Bar Chart

Or as a radar graph:

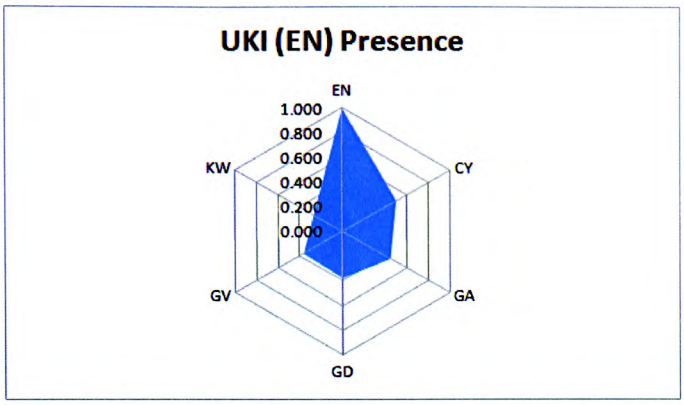


Figure 53 - S' (UKI) - Presence values – Radar Graph

The radar graph here is particularly instructive if we compare it to the previous radar graph for the EU major languages' presence as shown in Figure 44 above:

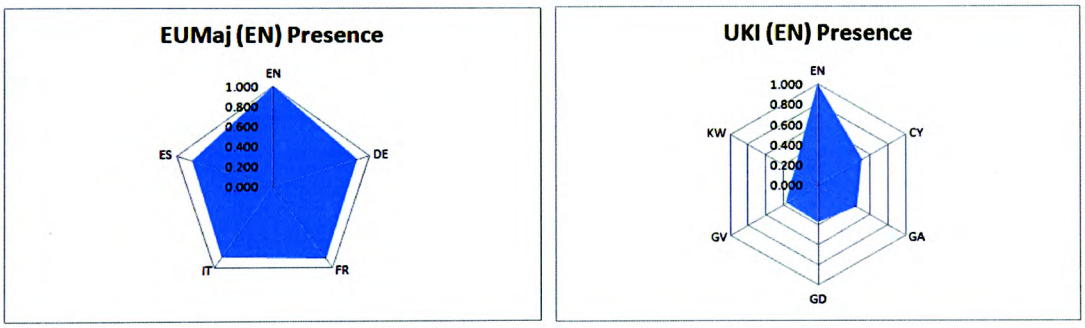


Figure 54 - S'' (EUMaj) and S' (UKI) Comparison

These two graphs give a good picture of a 'good' degree of presence within a constellation (S'' EU Maj) with a poorer one (S' UKI). If the UKI's presence were the same as for the EU major languages, it could be concluded that all or most of the languages within that constellation were in a healthy state. But, as it is, the radar graph of S' UKI shows clearly that English is the dominant language within that constellation, and that, to varying degrees, the Celtic languages that make up the rest of that constellation have relatively weak presences.

This is even more clearly illustrated by the following bubble graph:

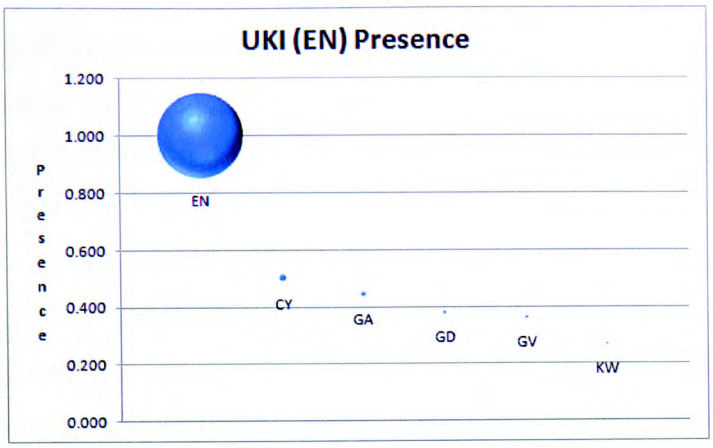


Figure 55 - S' (UKI) - Presence values – Bubble Graph

Here, it is quite graphically shown just how weak the presence of the Celtic languages are with constellation S' UKI. Indeed, if we push the analogy to a heavenly constellation, we could barely consider the Celtic languages as moons orbiting a planet; more like asteroids. To compare what a healthy constellation looks like we can show side by side Figure 55 with Figure 45 above, showing the presence values for the EU major languages:

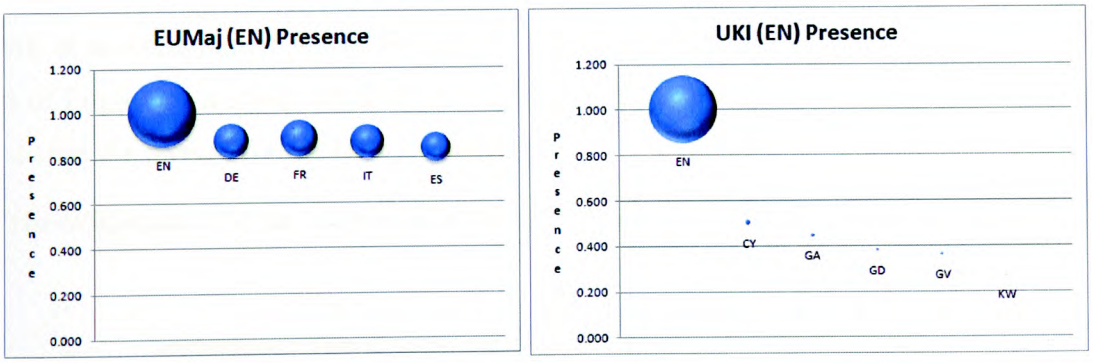


Figure 56 - S'' (EUMaj) and S' (UKI) Comparison

In all honesty, this is not an unexpected result, and tends to confirm what most researchers in the field would already know: that the Celtic languages of the UK and Ireland are dwarfed by use of English within those two states. Again, this would reinforce the validity of the results obtained from WkScrape, by confirming a picture that we already know.

One might argue that it is unfair to compare the situation of English, which has its principal strength not in the United Kingdom, but in the United States and elsewhere, with languages that have very little use outside of the United Kingdom. To a certain extent this is a valid point, but it must be remembered that the web does not have any physical boundaries and that, in the on-line world at least, Welsh, Irish, Gaelic and the other Celtic languages compete with English for the attention of persons inhabiting constellation S' UKI. On this basis, it is believed that this is a valid comparison of presence. While the majority of human linguistic activities do not occur online, increasingly the online world and the web in particular are becoming important areas of communication and therefore linguistic competition. If linguistic presence on the web is not healthy, then that may either impact on activity in the real world, or alternatively, may be a symptom of weak offline activity.

4. S' – Germany and the Netherlands

There is little point in repeating the same analysis as above for the somewhat artificial constellation of Germany and the Netherlands, which we will style S' (GEN). Germany and the Netherlands share much in common, and in the border regions there are undoubtedly some points of contact. But it is somewhat strained to consider these two nations as one linguistic constellation. The reasons for doing so are to simplify an analysis and to allow the three very different dialects of Frisian to be analysed together. The situation for the Frisian languages is similar to that of the Celtic languages in the UKI constellation. The purpose of this analysis is to look at a constellation where English is not the central language. It is possible that the status of English as a supercentral or even hypercentral language within the EU and the world may distort the picture for the UKI constellation.

The Wikiscores for the languages in the GEN are as follows:

Language	Raw 1	Pop 1	Raw 2	Pop 2
DE	13,785.825	104.359	8,507.685	64.403
NL	7,420.822	323.911	4,211.913	183.845
FY	139.197	266.490	61.978	118.657
FRR	7.493	802.854	2.712	290.521
STQ	17.396	5,798.711	6.124	2,041.209

Table 46 - S' (GEN) – Wikiscores

From these scores we can generate a number of different presence values by using different languages as the base. If we use English, German and Dutch as the base, we can calculate the following different presence values:

Base EN

Language	Raw1	Raw2
DE	0.875	0.846
NL	0.818	0.780
FY	0.453	0.386
FRR	0.185	0.093
STQ	0.262	0.169

Base DE

Language	Raw1	Raw2
DE	1.000	1.000
NL	0.935	0.922
FY	0.518	0.456
FRR	0.211	0.110
STQ	0.300	0.200

Base NL

Language	Raw1	Raw2
DE	1.069	1.084
NL	1.000	1.000
FY	0.554	0.494
FRR	0.226	0.120
STQ	0.320	0.217

Table 47 - S' (GEN) Different Presence Values (EN - DE- NL)

As can be expected, by using languages that have lower Wikiscores as the base for the presence formula, we increase the presence values for all the languages within this constellation. Thus, a change from base EN to base NL increases the presence value of West Frisian from .453 to .554, an increase of some 22.2%.

Thus there are a number of different presence values that a language can have depending on the analysis. If we look at West Frisian from the perspective of its role and status solely within the context of the Netherlands, we arrive at one value. If we look at how it is faring within a wider context of Germanic languages in the general region, we arrive at another value. If we look at its presence on the web from the point of view of its place within the wider European Union context we can derive yet another value.

While the presence values for all five languages within this constellation change if we change the base language, a graphing of the differences does not show massive differences.

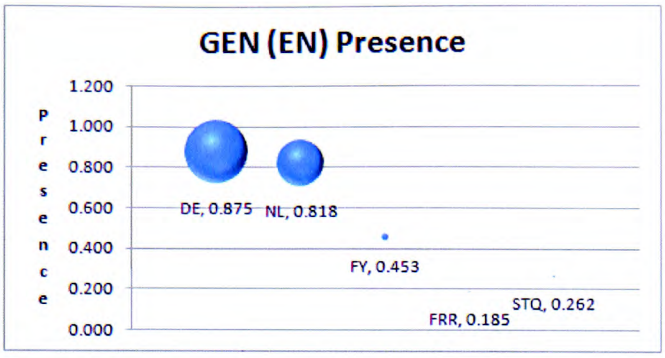


Figure 57 - S' (GEN) - Presence Values (Base EN)

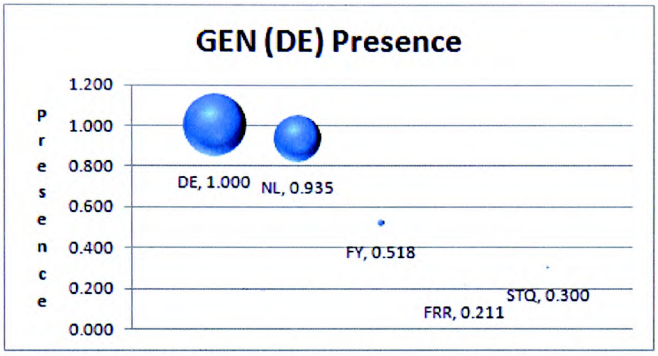


Figure 58 - S' (GEN) - Presence Values (Base DE)

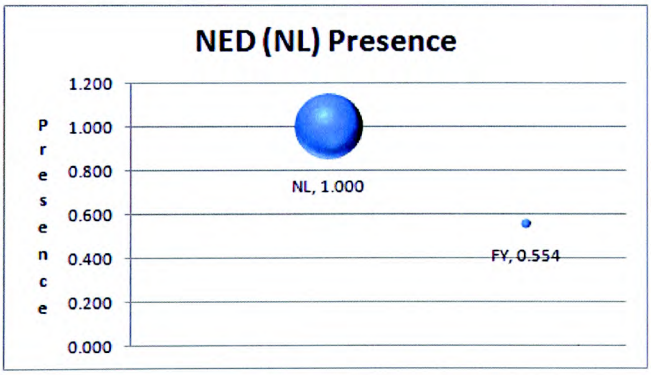


Figure 59 - S' (NED) - Presence Values (Base NL)

It is hard to discern the changes from the above graphs, but the following graph illustrates well how slight the differences are if we change the base language for presence calculation:

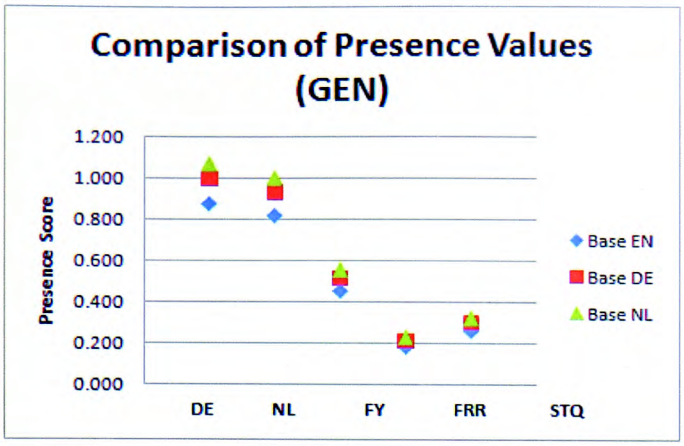


Figure 60 - S' (GEN) - Presence Values Differences (EN DE NL)

It is difficult to determine what value should be preferred. On the one hand, West Frisian speakers will normally be fully competent in Dutch and would have full access to the Dutch language edition of Wikipedia. Likewise, many Frisians have competence in German and English and may have access to those language editions as well. In the world before the web, we could be comfortable analysing West Frisian solely within its normal sphere in the Netherlands constellation, but the web changes the way languages interact and compete and we can no longer look at such a language from solely one perspective. This point is lost if we look at the Celtic languages of the UK and Ireland since their perspective is almost exclusively in comparison to English.

5. S'' – All Target Languages

So far we have looked at the major languages of Europe, as well as the UKI and GEN constellations. In would be instructive to look at a number of other constellations, in particular Spain, Germany, France and Italy, which are constellations that are broadly comparable to UKI, however space prevents this. At this point we will simply present the results for all the twenty languages that have been analysed as part of this study. The goal for this section will be principally to develop the language classification system that was presented in on page 94 above.

The following table gives the Wikiscores for all twenty languages with constellation S'' that have been scraped and analysed as part of this study. The goal is to produce a fair sampling of the languages of Western Europe, with an ultimate goal of analysing all the languages currently spoken and used with Western Europe.

Language	Raw 1	Pop 1	Raw 2	Pop 2
EN	53,608.917	66.021	44,392.297	54.670
DE	13,785.825	104.359	8,507.685	64.403
FR	15,697.135	88.385	9,398.627	52.920
IT	12,967.499	202.934	7,315.987	114.491
ES	9,672.214	23.288	6,431.845	15.486
CY	234.429	337.213	84.251	121.190
GA	126.314	297.332	59.800	140.765
GD	61.599	939.106	18.534	282.560
GV	50.708	1,906.806	18.247	686.163
KW	18.275	4,249.979	4.033	937.995
NL	7,420.822	323.911	4,211.913	183.845
FY	139.197	266.490	61.978	118.657
FRR	7.493	802.854	2.712	290.521
STQ	17.396	5,798.711	6.124	2,041.209
DA	1,438.980	254.687	738.519	130.711
IS	235.126	829.856	91.458	322.792
BR	280.341	754.250	107.228	288.495
CA	3,835.679	353.734	2,003.373	184.755
GL	708.751	208.763	358.611	105.629
SC	17.640	12.467	5.292	3.740

Table 48 - S'' (EU All) – Wikiscores

From these Wikiscores we can derive the presence values as follows:

Language	Raw1	Raw2
EN	1.000	1.000
DE	0.875	0.846
FR	0.887	0.855
IT	0.870	0.832
ES	0.843	0.819
CY	0.501	0.414
GA	0.444	0.382
GD	0.378	0.273
GV	0.361	0.271
KW	0.267	0.130
NL	0.818	0.780
FY	0.453	0.386
FRR	0.185	0.093
STQ	0.262	0.169
DA	0.668	0.617
IS	0.501	0.422
BR	0.518	0.437
CA	0.758	0.710
GL	0.603	0.550
SC	0.264	0.156

Table 49 - S'' (EU All) – Presence Values

We can visualise the data by making use of two bubble diagrams, the first of which graphs the predicted Tier I to Tier IV languages; the second, the Tier V to Tier VI languages:

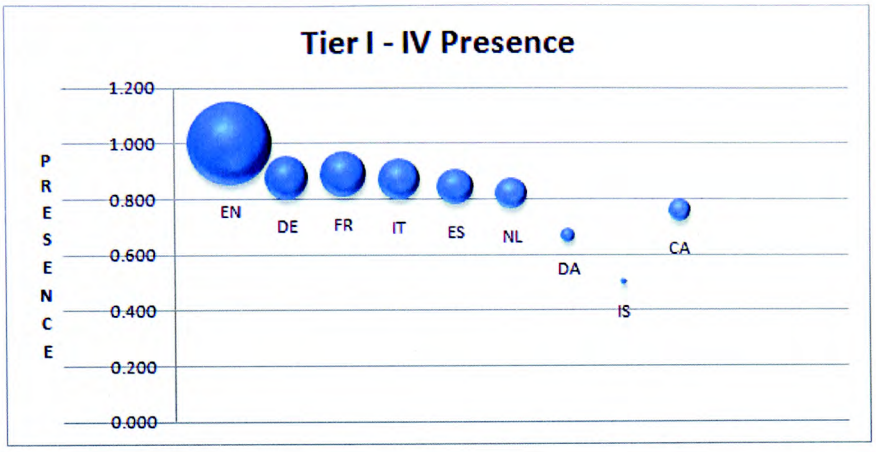


Figure 61 - Tier I to Tier IV – Presence Values

Figure 61 shows all of the predicted Tiers I to IV languages, but includes Catalan, which was predicted to be a Tier V language, because its presence was clearly greater than that for Danish and Icelandic, which were predicted to be Tier IV languages. This chart, in general, shows that, but for Icelandic, there is generally a good presence for most of the predicted Tier I to IV languages (and for Catalan).

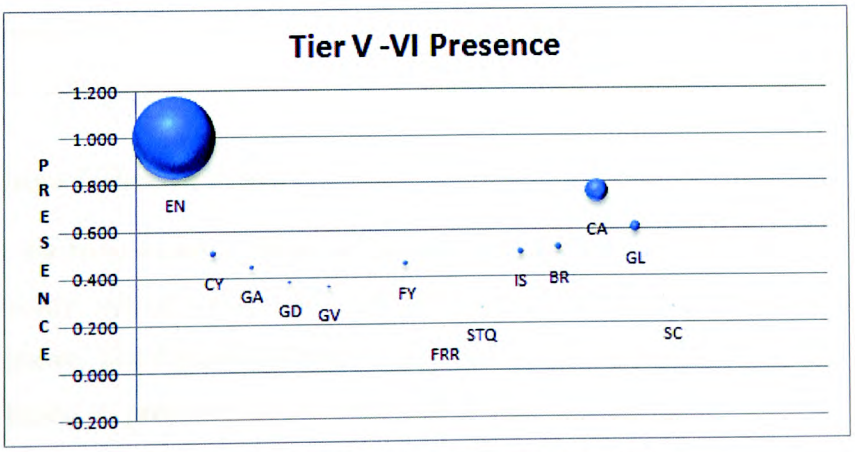


Figure 62 - Tier V to Tier VI – Presence Values

Figure 62 shows the presence values for the minority and regional languages predicted to be in Tiers V to VI, with the addition of Icelandic. The results are generally within a similar range, with a few notable exceptions. Firstly, Catalan is outperforming the other regional languages, and, as is shown in Figure 61, it even outperforms Danish. This will not be a surprise to those who are familiar with the situation of Catalan.

But the chief value of Figure 61 and Figure 62 is to graphically illustrate exactly how the various language editions of Wikipedia compare to each other. We have two distinct methods of analysing such data, and each gives a slightly different picture. If we look at both the Wikiscores and the presence values, we see a marked difference between the values arrived at for a majority language as compared to a minority language. It is possible to visually understand just how differently the majority and minority languages are faring on the web.

If we were to draw some preliminary conclusions in line with Crystal's prediction that a large number of languages may soon face extinction in the coming centuries, we could possibly use the presence values as an indicator. It seems obvious that a presence value of at least .600 gives at least some indication that the language presence is healthy, but that perhaps a value of at least .800 is needed to show some real health. If this prediction were to prove accurate, then none of the languages other than the major languages and Catalan may look at long term survival.

Of course, this may be a premature prediction and much study over time and comparison with real world results would be needed before such ominous conclusions could be drawn. But, it is submitted that by providing these indicators, the WkScrape program and the method of analysing the data proposed give a good basis for the examination of these issues.

B. Language Classification

Now that we have calculated the presence values for the twenty languages that are the target of this study, we can now revisit the language classification scheme first proposed in Section IV.B above. The following table quickly summarises the predicted results of language classification based on presence on the web (only the target languages are shown):

Tier	Example Languages	Expected Score
I	English	1
II	Spanish, French, German	.900 - .999
III	Italian, Dutch,	.800 - .899
IV	Danish, Icelandic	.700 - .799
V	Welsh, Irish, Scots Gaelic, Manx, Catalan, Frisian, Sardinian, Galician	.600 – .699
VI	Breton, Cornish, North Frisian, Saterland Frisian	.500 - .599
VII		.400 - .499
VIII		.300 - .399
IX		.200 - .299
X		<.199

Table 50 - Language Classification – Predicted

Tier	Example Languages	Actual Score
I	English	1
II		.900 - .999
III	German, French, Italian, Spanish, Dutch	.800 - .899
IV	Catalan	.700 - .799
V	Danish, Galician	.600 – .699
VI	Welsh, Breton, Icelandic	.500 - .599
VII	Irish, West Frisian	.400 - .499
VIII	Manx, Scots Gaelic	.300 - .399
IX	Cornish, Saterland Frisian, Sardinian	.200 - .299
X	North Frisian	<.199

Table 51 - Language Classification – Actual

The actual prediction exercise was not successful, with only four languages correctly predicted: Italian, Dutch, Breton and Galician. The following languages were overestimated with respect to their actual presence values: Spanish, French, German, Danish, Icelandic, Welsh, Irish, Scots Gaelic, Manx, Frisian, Sardinian, Cornish, North Frisian and Saterland Frisian. Only one language was underestimated: Catalan. Overall, there was a tendency to believe that a language had a greater presence than it really had.

If we look at the table, there is a fairly interesting spread among the tiers. But it is clear that the major languages of Western Europe cannot be classed in Tier II. That Danish and Icelandic are 'underperforming' is not particularly surprising given the relatively smaller populations (when compared to German and French, for example). Likewise the result for

Catalan is hardly surprising in that it comes only slightly below Dutch, which has half the number of speakers as Catalan.

What is surprising, perhaps, is that there no language is currently classed in Tier II. This clearly shows just how dominant English is in the world today. English has had an enormous head-start on the web, and is the language of the United States, not to mention a number of other large, industrialised nations. Probably no language can, at the moment, challenge English in this regard. Tier II is perhaps best seen as a tier reserved for only a few of the very large languages, perhaps: Chinese, Hindi, Spanish or Arabic, but these were not target languages of this study. It is possible that these languages may have the potential to generate sufficient momentum that they can come close to English in terms of presence. Therefore, of the languages studied, it is perhaps only Spanish that is really underperforming, since it potentially has the speaker numbers and the supercentral position in a large part of the world to possibly, in the future, become a Tier II language. It would be interesting to be able to study the other candidate Tier II languages (Chinese, Hindi and Arabic) to see where they stand at the moment, and to chart their growth over the coming years.

What does seem clear from this classification scheme is that numbers of speakers appears to be the primary driver behind presence values. That French, German and Italian on the one hand and Welsh, Breton and Icelandic on the other, have roughly the same presence may be caused by the fact that they have roughly the same populations. This would account for their greater presence than, say, Irish or Gaelic, which have much fewer speakers. Thus, it seems that some more work may need to be done to provide a more numbers based approach to presence.

In general, the language classification system as proposed appears to be a workable methodology. Since we are measuring language presence on-line we can look at each language as having a discrete presence on the web and therefore having a value comparable with other languages. On the web, for the most part, all information is available to all users and, in this regard, the web is a great equaliser.

This language classification scheme is not intended solely to be an indicator of presence for Wikipedia only, but rather is designed to analysis all aspects of language presence. It must be remembered that Wikipedia is a user-generated Web 2.0 application that does not receive state support. There are other areas of the web where this is not true. Since Wikipedia is user-

generated, it is understandable that the number of users that a community can muster would be an important factor in the size of the application. In other areas of the web that is not true. Language presence ultimately is a very complicated picture and one aspect: state support of a language, can be underestimated.

The validity of the proposed language classification scheme awaits further testing and analysis.

C. Conclusions Regarding Presence

A goal of this study has been to propose, develop and demonstrate a method of calculating presence values for various languages by using the data supplied by the WkScrape program and from that data calculating Wikiscores and then presence values. As was demonstrated above, Wikipedia provides its own methods of ranking the different language editions of Wikipedia, one of which is to rank the different editions by article count. If this method is accurate, then there would be no need to obtain the data from WkScrape and we could calculate presence from article counts alone. However, as was shown, Wikipedia's use of article counts does not necessarily give the full picture.

The following table shows, in column 2, the article counts obtained from Wikipedia for November 2010 together with, in column 3, a calculation of presence, using the same formula as provided above in Section IV.D.5 above. For reference, the presence values derived from WkScrape data are provided in columns 4 and 5.

As can be seen there are noticeable differences between the two sets of calculations, especially if we look at the values for the minority languages. Cornish, for example, obtains a presence score of .504 if we look at article counts, whereas the data obtained from WkScrape would give it a very significantly lower presence value of .267. The reasons for these

	Article Count		Wikiscore	
Language	Num Articles	Presence	Presence 1	Presence 2
EN	3,500,000	1.000	1.000	1.000
DE	1,200,000	0.929	0.875	0.846
FR	1,000,000	0.917	0.887	0.855
IT	758,000	0.898	0.870	0.832
ES	676,000	0.891	0.843	0.819
CY	30,000	0.684	0.501	0.414
GA	12,000	0.623	0.444	0.382
GD	8,100	0.597	0.378	0.273
GV	3,600	0.543	0.361	0.271
KW	2,000	0.504	0.267	0.130
BR	36,000	0.696	0.818	0.780
NL	656,000	0.889	0.453	0.386
FY	18,000	0.650	0.185	0.093
FRR	1,000	0.458	0.262	0.169
STQ	2,000	0.504	0.668	0.617
DA	139,000	0.786	0.501	0.422
IS	30,000	0.684	0.518	0.437
CA	296,000	0.836	0.758	0.710
GL	65,000	0.735	0.603	0.550
SC	2,500	0.519	0.264	0.156

Table 52 - Comparison of Presence Calculations

differences are related to the ability of WkScrape to obtain more information about each article and to combine this into a more accurate calculation of the actual information available in the different language editions of Wikipedia.

If we graph the two different sets of presence values we can see that the differences for the five major European languages are minor:

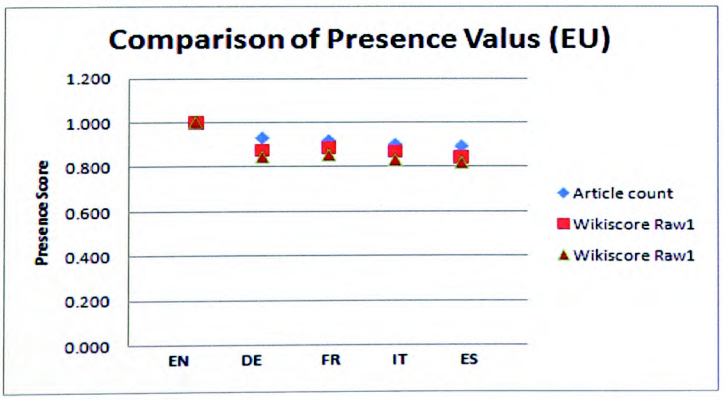


Figure 63 - Comparison of Presence Values (EU)

While a graph of the minority languages shows very markedly different results:

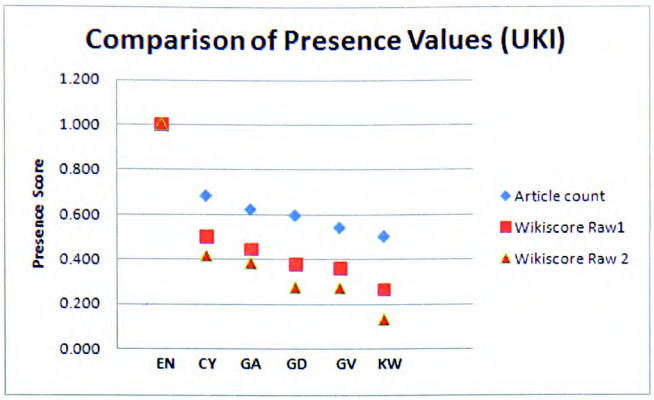


Figure 64 - Comparison of Presence Values (UKI)

WkScrape does not change the overall picture for majority languages, but gives a more accurate and much different picture for minority languages.

VI. Future Work and Conclusions

A. Lessons Learned and Further Work on WkScrape

The number of bytes that each article uses is not a reliable indicator of the information contained in that article. Any future study that attempts to estimate language use on the web by looking at the computing data needs to be clear that computer data is not always read by humans and much of it is the overhead code that, while necessary to present the data, is not useful to a user. If a method could be found to separate the content data from the underlying code, then such a method could be valid. In fact the WkScrape program does that.

The production of 'population' statistics whereby the size of a language edition of Wikipedia is factored by the estimated number of speakers produces unreliable, and sometimes absurd, results and should be rejected for languages that have small populations. This is amply demonstrated by the results returned for the two 'revived' Celtic languages: Cornish and Manx, which have extremely small numbers of speakers.

A number of raw data indicators proved to be more trouble to collect than they were worth. There does not seem to be any net difference to collecting and analysing internal links and external links.

There seems to be a direct correlation between the number of words in an article and the number of images. This correlation is not surprising given that those articles that have been well worked tend to create or collect images and other visuals to support their work. Some thought therefore could be given to whether it was a useful exercise to collect this data separately. In the event a direct correlation could be shown, which is a conclusion is borne out by this study, then it would save some processing time in not analysing a Wikipedia article for images. The same could be said about the number of links. Any further study could look at these correlations to see if time and energy could be saved by solely looking at the number of words in an article.

Much time and effort was spent attempting to weight the number of words for the various languages. Ultimately, the use of the Omniglot versions of the Bible was not satisfactory. In any event, the use of weightings did not have a great impact on the overall results and some time and effort could be saved by not carrying out this exercise.

While the sample rate of 3000 articles per analysis seems to give work results, some thought should be given to sampling larger datasets. While this would not necessarily increase the accuracy of the results, that can only be determined after such analyses have been carried out.

There are still some small errors and inefficiencies in the WkScrape program itself and more development time is needed. In particular, the program needs to be more efficient in its scraping of the index for each language and should be made more efficient in the number of requests it makes to Wikipedia's servers.

The range of languages chosen was incomplete. Since all the languages of Europe use only three basic alphabets: Roman, Greek and Cyrillic, there is no reason that the study cannot be extended to all the languages of Europe.

One notable feature not explored was development over time. Indeed, the chief benefit of this study would lie in showing how the various languages may be improving or degrading over time. Of course, a proper study of this may take years, if not decades, before any real trends can be discerned and conclusions drawn.

Consideration should be given to adding further criteria to the Wikiscores. For example, each article was counted as a discrete unit. However, an encyclopaedia, such as Wikipedia, needs to have good range of subject areas in addition to a large number of articles. Thus, some thought could be given to looking into master subject areas, such as 'science', 'literature', 'sports', etc. to see whether than has any impact on presence determination.

B. Future Work

As was stated in the Research Methodologies chapter, the purpose of an EDA analysis is to draw a picture from the data that can then be used to form working hypotheses for future study. Some potential hypotheses are:

The tests were carried out in 2010. Further tests could be carried out in the future, which could reveal trends and reveal other potential lines of research.

English is clearly showing a tendency to be, at least as far as European languages are concerned, of a different order of magnitude than the other majority languages of Europe (German, French, Italian and Spanish). The English language edition of Wikipedia is nearly 3.5 times as large as the French edition, the next largest. Previous studies have led us to

expect this result. Will this situation prevail? Will the other majority languages 'close the gap' or conversely will English continue to be on a different scale?

Italian is a very high performing language, especially when compared to German, which has nearly the double the speakers. A number of questions are raised by this fact. Are Italian speakers more literate, or do they have a greater appreciation for Wikipedia? Is this down to a few individuals or is it a community effort? Are some languages more inclined to favour depth over breadth? Are other communities more content to have 'thin' content spread more broadly?

Spanish conversely is a very underperforming language. This confirms findings from other researchers. The cause of this merits further attention. If we look at the situation of Italian, which is a related romance language, we see two extremes. Why is one community more likely to produce material than another? What are the reasons for this difference? Are they economic, or are there other factors?

Moving to the presence values we can see that the values of all the major languages were, however, roughly comparable, the range being .877 (French) to .843 (Spanish). Thus, while English, as the base, scores 1.000, the other supercentral/large central languages of Europe fall within the .8 decile on the presence score. This is perhaps a good indicator of where a central or supercentral language should score. An interesting test would be to see where the other supercentral/large central languages (e.g. Chinese, Arabic, Japanese) would score.

The results for the minority languages studied were not surprising to anyone acquainted with the literature or the actual situation of these languages. However, we can, at least with respect to the presence on Wikipedia, start to see some detail, and have the possibility of making comparisons. Therefore, the calculations for the presence of each language provide a useful tool in measuring and comparing the content of each language edition of Wikipedia, and may yet prove to be a useful tool in measuring other aspects of minority language use on the web. Comparisons can be made between the various minority languages, in terms of population, status and other factors. What are the causes of the differences between the languages? As an example, Breton has fewer speakers than Welsh, yet the presence values on Wikipedia are similar.

The language tier classification system, as refined, is a useful way to compare languages. At the moment, at least in respect to the target languages studied, all of which are spoken in

industrialised Western Europe, the correlation between numbers of speakers and tier ranking seems to be clear. However, if more languages were studied, especially those of less developed nations, a more nuanced picture might appear. Thus, the language classification system as proposed herein should only be considered as a preliminary proposal for the classification of languages on the web. Further thought may be given to refining, expanding and testing it on other data sets.

The ten-tiered language classification system for on-line language measurement produced some interesting results. By running the tests and then assigning the languages to the various tiers, a pattern emerged. It turned out that the initial predictions tended to be on the high side, and this is mainly due to the extremely large size of the English edition of Wikipedia. Thought could be given to using another language as the base point of comparison, which might yield more nuanced results. However, the point seems to be validated that we can *a priori* guess with some accuracy where a language will 'rank' by looking at the factors enunciated in the classification table.

More importantly, the concept of presence needs to be tested in other areas of the web besides Wikipedia. The general notion of presence in Wikipedia is only suggested as one method of measuring presence. It is possible that Wikipedia may provide a fairly accurate 'proxy' indicator of presence in that in forms a perfect test case for measuring presence on the web, but this can only be determined if further analyses of presence are made in other areas of the web. Suitable candidates for this study could look at governmental sites, other Web 2.0 applications, news organisations, other media, etc. However, it may transpire that Wikipedia is a unique Web 2.0 phenomenon that is prone to exaggerate or favour certain factors and languages. Only further tests on other parts of the web can confirm or deny this.

It may be possible to adapt the general methodology elaborated in this study to other areas of the web. Certainly the breadth and depth approach is broadly applicable to all most websites. Furthermore the language constellation system is a useful way of grouping languages and provides language communities with a model framework for comparison.

C. Conclusions

The WkScrape program produces reliable data that is both confirmed by and confirms previous studies. Wikiscores are generally reliable indicators of the size of a particular language edition of Wikipedia. In general, it is useful to point out that the results of the data produced by WkScrape would probably not come as a surprise to anyone familiar with the language situation in western Europe. Indeed, but for a certain feeling that some numbers for the minority languages may be on the low side, the results are almost entirely within the margins of expected results for these languages. Thus, with respect to majority languages, while there is nothing necessarily new in the presentation of the results of this study, the WkScrape program and the methodology that it uses produce results that, at least with respect to majority languages, accord with prior work in the field.

The originality and chief value of this study lies in the extension of the methodology to a range of minority languages. No other previous study has looked at minority languages on this scale and using the same methodology as applied to majority languages. With the exception of the Mas study in 2003, which used a 'search engine' approach and which is impossible to use further, since the search engine he used no longer exists, no other study has looked at minority languages alongside majority languages.

Using a population analysis, the Italian edition of Wikipedia is almost double that of German, and over three times higher than English. This is not a fact that is borne out by Wikipedia's measure of using number of articles, and is a proof that the 'presence' measure used in this study gives a fuller and arguably more nuanced picture of language material available on the web. Thus, this study has shown that the use of article count alone is not necessarily the best measure of the size and value of a particular language edition of Wikipedia. The methodology employed in this study provides a more complete and more accurate picture.

The data visualisation techniques used in this study have a useful role in visually demonstrating the nature of the problem that minority languages face. The following diagram rather graphically illustrates just how 'peripheral' the minority languages are in Europe:

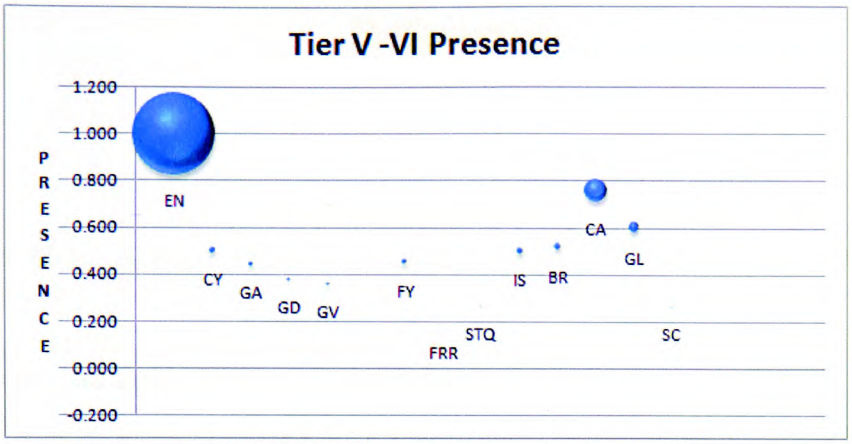


Figure 63 - Tier V to Tier VI – Presence Values

As was pointed out in the Introduction, it is possible that many of these languages may well disappear within the immediate future. If we are to study minority language use it is useful to be able to make comparisons and to give to interested parties accurate and useful data.

References

Abley, M. (2004). *Spoken Here : Travels among Threatened Languages.* London: Heinemann.

Babel Project. (1997). Web Languages Hit Parade Retrieved February 2007, from http://alis.isoc.org/palmares.en.html#liste_langues

Baeza-Yates, R., & Castillo, C. (2000). *Caracterizando la web chilena.* Encuentro chileno de ciencias de la computacion. Sociedad Chilena de Ciencias de la Computacíon. Punta Arenas, Chile.

Baeza-Yates, R., & Castillo, C. (2002). Caracterizando la Web chilena Retrieved December 2006, from mazinger.sisib.uchile.cl/repositorio/pa//b20028231220caracterizandolawebchilena.doc

Baeza-Yates, R., & Castillo, C. (2004). *Caracterısticas de la web chilena 2004.* Center for Web Research, University of Chile, 2005.

Baeza-Yates, R., Castillo, C., & López, V. (2006). Characteristics of the Web of Spain. *Cybermetrics, 9*(1).

Bergman, M. K. (2001). White Paper:The Deep Web: Surfacing Hidden Value. *The Journal of Electronic Publishing, 7*(1).

Blumenstock, J. E. (2008). *Size Matters: Word Count as a Measure of Quality on Wikipedia.* Paper presented at the WWW 2008, Beijing, China.

Bolshakov, I. A., & Galicia-Haro, S. N. (2003). Can we correctly estimate the total number of pages in Google for a specific language? *Computational Linguistics and Intelligent Text Processing. 4th International Conference, CICLing 2003. Proceedings (Lecture Notes in Computer Science Vol.2588). Springer-Verlag. 2003*, 415-419.

Brenzinger, M., A. Yamamoto, N. Aikawa, D. Koundiouba, A.Minasyan, A. Dwyer, . . . Zepeda, O. (2003). Language vitality and endangerment. Retrieved from http://www.unesco.org/culture/en/endangeredlanguages.

Cafarella, M. J., & Etzioni, O. (2005). A search engine for natural language applications. *International World Wide Web Conference Committee (IW3C2)* Retrieved December 2006, from www.cs.washington.edu/homes/etzioni/papers/be_www2005.pdf

Crovella, M., & Krishnamurthy, B. (2006). *Internet Measurement: Infrastructure, Traffic and Applications*: Wiley.

Crystal, D. (1997). *English as a Global Language*: Cambridge University Press.

Crystal, D. (2001). *Language and the Internet*: Cambridge University Press.

Crystal, D. (2002). *Language Death.* Cambridge: Cambridge University Press.

Crystal, D. (2004). *The Language Revolution.* Cambridge: Polity Press Ltd.

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics* (6th ed.). Oxford.

References

Cunliffe, D. (2004). Promoting minority language use on bilingual Web sites. *Mercator Media Forum, 7*, 42-53.

Cunliffe, D. (2005). Party political web sites and minority languages: some initial observations from Wales. In F. Sudweeks, H. Hrachovec & C. Ess (Eds.), *Proceedings of the 5th international conference on Cultural Attitudes towards Technology and Communication (CATaC 2005)* (Vol. 696–701). Tartu, Estonia.

Cunliffe, D., & Harries, R. (2005). Promoting minority-language use in a bilingual online community. *The New Review of Hypermedia and Multimedia, 11*(2), 157–179.

Cunliffe, D., & Roberts-Young, D. (2005). Online design for bilingual civil society: a Welsh perspective. *Interacting with Computers, 17* (1), 85–104.

de Swaan, A. (2001). *Words of the World.* Cambridge, England: Polity Press.

Deere, A., & Cunliffe, D. (2005). *Bilingual websites in jurisdictions requiring minority language use: Effective implementation of policies and guidelines.* Paper presented at the 10th International Conference on Minority Languages, Trieste, Italy. http://www.slori.org/conference/pagina.php?pag=n&dett=8&id=26

Deere, A., & Cunliffe, D. (2009). Bilingual Websites in Jurisdictions Requiring Minority Language Use: Effective Implementation of Guidelines. In S. Pertot, T. M. S. Priestly & C. Williams (Eds.), *Rights, Promotion and Integration Issues for Minority Languages in Europe.* Basingstoke: Palgrave MacMillan.

Efthimiadis, E., & Castillo, C. (2004). Charting the Greek Web. *Proceedings of the Conference of the American Society for Information Science and Technology (ASIST), Providence, Rhode Island, USA, November 2004.*

Egan, K. (2000). *Bilingual Website design: A study and report on Welsh/English bilingual Websites.* MSc, University of Glamorgan, Pontypridd.

Ethnologue.com. (2008). Ethnologue, Languages of the World (website) Retrieved March 2008, from http://www.ethnologue.com/

European Commission. (2006). *Special Eurobarometer 243: Europeans and their Languages.* Retrieved from http://ec.europa.eu/public_opinion/archives/ebs/ebs_243_en.pdf.

Evas, J. (2001). *Snapshot Survey: Websites of Organisations Complying with Statutory Language Schemes.* Aberdare, Wales: Linguacambria Cyf.

Fantognan, X. (2005). A note on African languages on the internet. In U. I. f. Statistics (Ed.), *Measuring linguistic diversity on the internet.* Paris: United Nations Educational, Scientific and Cultural Organisation.

Fishman, J. A. (1991). *Reversing Language Shift: Theoretical and Empirical Foundations of Assistance to Threatened Languages* (Vol. 76). Clevedon: Multilingual Matters.

Fishman, J. A. (2000). *Can Threatened Languages Be Saved?: Reversing Language Shift, Revisited - A 21st Century Perspective.* Clevedon: Multilingual Matters.

Fishman, J. A. (2002). Endangered minority languages: prospects for sociolinguistic research. *International Journal on Multicultural Societies, 4*(2), 270-275.

FUNREDES. (2005). Lenguas y Culturas en la Red, 2005 Retrieved 2007, from http://funredes.org/lc/english/medidas/sintesis.htm

References

Gerrand, P. (2007). Estimating linguistic diversity on the Internet: A taxonomy to avoid pitfalls and paradoxes. *Journal of Computer-Mediated Communication, 12*(4), Article 8.

Gomer, R. (2003). *Websites survey: websites of organisations that have a statutory Welsh Language Scheme*. Wales: Cwmni Cymad & Bwrdd yr Iaith Gymraeg - Welsh Language Board.

Gomes, D., & Silva., M. J. (2003). A characterization of the portuguese web. *Proceedings of 3rd ECDL Workshop on Web Archives, Trondheim, Norway, 2003.*

Greenberg, J. H. (1956). The Measurement of Linguistic Diversity. *Language Problems and Language Planning, 32*(1), 109-115.

Grefenstette, G., & Nioche, J. (2000). Estimation of English and non-English Language Use on the WWW Retrieved March 2007, from http://citeseer.ist.psu.edu/cache/papers/cs/20658/http:zSzzSzwww.jnioche.freesurf.frz SzpaperszSzL76grefen.pdf/grefenstette00estimation.pdf

Grenoble, L. A., & Whaley, L. J. (2006). *Saving Languages: An Introduction to Language Revitalization*: Cambridge University Press.

Guinovart, X. G. (2003). A lingua galega en Internet. In A. Bringas & B. Martín (Eds.), *Nacionalismo e gobalización: lingua, cultura e identidade* (pp. 71-88). Vigo: Servicio de Publicacións da Universidade de Vigo.

Gulli, A., & Signorini, A. (2005). The indexable Web is more than 11.5 billion pages *Poster proceedings of the 14th international conference on World Wide Web, Chiba, Japan, 2005* (pp. 902–903). Chiba, Japan: ACM Press.

Harries, R. (2003). *Developing a bilingual online community of practice.* MSc, University of Glamorgan (MSc Thesis), Pontypridd.

Harries, R., & Cunliffe, D. (2004). Welsh use in a bilingual online community: An initial analysis of Pen i Ben. In F. Sudweeks & C. Ess (Eds.), *Proceedings of the 4th international conference on Cultural Attitudes towards Technology and Communication (CATaC 2004)* (pp. 389–392). Karlstad, Sweden.

Honeycutt, C., & Cunliffe, D. (2010). The Use of the Welsh Language on Facebook: An initial investigation *Information, Communication & Society, 13*(2), 226 - 248.

Janse, M., & Tol, S. (2003). *Language Death and Language Maintenance: Theoretical, Practical and Descriptive Approaches*. Amsterdam; Philadelphia: John Benjamins Publishing Company.

Jarvis, M. (2000). *Bilingual Web site design: A study and recommendations.* MSc Thesis, University of Glamorgan, Pontypridd.

Jones, C. O. (2009). Look It Up on Wicipedia. *Planet* (195), 27-31.

Kelly-Holmes, H. (2006). Irish on the World Wide Web: Searches and sites. *Journal of Language and Politics, 5*(2), 217-238.

Large, A. (2002). The New Babel: Language Barriers on the World Wide Web. *Journal of Universal Language, 3*, 77-95.

Large, A., & Moukdad, H. (2000). Multilingual Access to Web Resources: An Overview. *Program, 34*(1), 43-58.

References

Lavoie, B. F., & O'Neill, E. T. (1999). How "World Wide" Is the Web? Trends in the Internationalization of Web Sites. *Annual Review of OCLC Research 1999* Retrieved 7/29/2007,from http://digitalarchive.oclc.org/da/ViewObjectMain.jsp?fileid=0000002655:000000059202&reqid=21527&frame=false

Lewis, M. P., & Simons, G. F. (2009). *Assessing Endagerment: Maximizing Fishman's GIDS.* Retrieved from http://www.sil.org/~simonsg/preprint/EGIDS.pdf

Lim, E.-P., Vuong, B.-Q., Lauw, H. W., & Sun, A. (2006). Measuring Qualities of Articles Contributed by Online Communities. *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence.*

Mas i Hernàndez, J. (2003). La salut del català a Internet [The health of Catalan on the Internet]. Retrieved July 2007, from http://www.softcatala.org/articles/article26.htm

Mensching, G. (2000). The internet as a Rescue Tool of Endangered Languages: Sardinian Retrieved June 2007, from http://www.gaia.es/multilinguae/pdf/Guido.PDF

O'Neill, E. T., Lavoie, B. F., & Bennett, R. (2003). Trends in the Evolution of the Public Web. *D-Lib Magazine* Retrieved 29 May 2005, from http://www.dlib.org/dlib/april03/lavoie/04lavoie.html

Ó Néill, D. (2005). *Rebuilding the Celtic Languages: Reversing Language Shift in Celtic Countries.* Talybont, Ceredigion, Wales: Y Lolfa.

Ortega Soto, J. F. (2009). *Wikipedia: A quantitative analysis.* PhD Thesis, Universidad Rey Juan Carlos, Madrid. Retrieved from http://libresoft.es/Members/jfelipe/thesis-wkp-quantanalysis

Paolillo, J. C. (2005). Language Diversity on the Internet. In U. I. f. Statistics (Ed.), *Measuring linguistic diversity on the internet.* Paris: United Nations Educational, Scientific and Cultural Organisation.

Paolillo, J. C., & Das, A. (2006). Evaluating Language Statistics: The Ethnologue and Beyond. Report prepared for the UNESCO Institute for Statistics Retrieved July 2007, from http://ella.slis.indiana.edu/~paolillo/research/u_lg_rept.pdf

Paolillo, J. C., Pimienta, D., Prado, D., & al. (2005). *Measuring Linguistic Diversity on the Internet*: United Nations Educational, Scientific and Cultural Organization.

Pimienta, D. (2005). Linguistic diversity in cyberspace; models for development and measurement. In UNESCO Institute for Statistics (Ed.), *Measuring linguistic diversity on the internet.* Paris: United Nations Educational, Scientific and Cultural Organisation.

Pimienta, D., & Lamey, B. (2001). *Lengua española y culturas hispánicas en la Internet. Comparación con el inglés y el francés [The Spanish language and Hispanic cultures in the Internet. Comparison with English and French].* . Paper presented at the II Congreso Internacional de la Lengua Española, Valladolid, Spain.

Pimienta, D., Prada, D., & Blanco, D. (2009). *Twelve years of measuring linguistic diversity in the Internet: balance and perspectives.* Retrieved from http://unesdoc.unesco.org/images/0018/001870/187016e.pdf.

References

Richards, S. (2005). Internet Inequality in Wales, from http://www.wales-consumer.org.uk/research_policy/pdf/WCC47_Internet_Inequality_in_Wales_update_2006.pdf

Romero, D., & Vaquero, I. (1999). *Lingua e informatica: O galego na Rede*. Paper presented at the O galego ás portas do 2000, Vifo.

Thomson, R., & Cunliffe, D. (2005). Welsh identity on-line. In D. Cunliffe, R. Thomson & C. Williams (Eds.), *Proceedings of the first workshop on E-dentity: Borders and Identities in the Internet Age* (pp. 53-59). Treforest, Wales.

Tukey, J. W. (1977). *Exploratory Data Analysis*. New York: Addison-Wesley.

UNESCO Institute for Statistics. (2003). Measuring and monitoring the information and knowledge societies: a statistical challenge Retrieved July 2004, from http://portal.unesco.org/ci/en/files/12851/10711588385uis.pdf/uis.pdf

Van Belle, J.-P., Fellstad, R., Steele, C., & van Bakel, W. (2003). Multi-language websites in a multi-cultural country: a South African perspective *Third International Conference on Electronic Business (ICEB 2003)*. Singapore, 9-13 Dec 2003.

van Dijk, Z. (2009). Wikipedia and lesser-resourced languages. *Language Problems and Language Planning, 33*(3), 234-250.

Voß, J. (2005[a]). Measuring Wikipedia. In P. Ingwersen & B. Larsen (Eds.), *Issi 2005: Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics, Vols 1 and 2* (pp. 221-231). Stockholm: Karolinska Univ Press Ab.

Voß, J. (2005[b]). Measuring Wikipedia: empirical analysis of the free encylopaedia Retrieved 8 February 2008, from http://wm.sieheauch.de/files/ISSI2005-Measuring_Wikipedia-Presentation.pdf

Welsh Language Board. (2006). *Bilingual Software Standards and Guidelines, version 1.0.* Retrieved from http://www.bwrdd-yr-iaith.org.uk/.

Wyn Jones, E. (2007). *Survey of promoting technology in Welsh: Experimental project - Anglesey*. Retrieved from www.bwrdd-yr-iaith.org.uk/download.php/pID=82999.

Appendix A

SPSS Statistics for Samples for the Twenty Languages Studied

English

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	3000	3000	3000	3000	3000	3000
	Missing	0	0	0	0	0	0
Mean		53029.89	1375.03	948.41	4.27	184.85	10.64
Median		36133.50	719.00	300.00	1.00	116.50	5.00
Mode		20025[a]	255	0	1	75	2
Std. Deviation		50446.430	1959.695	1849.211	12.834	196.041	30.862
Range		681469	21373	20420	201	2398	1284
Minimum		18927	168	0	0	42	2
Maximum		700396	21541	20420	201	2440	1286
Sum		159089655	4125100	2845234	12814	554558	31912

Wiki Art Count	3,500,000
Population	812,000,000
Word weighting	1.000
Image weighting	50
Link weighting	2
Link weighting - External	10
WS-Raw1	53,608.917
WS-Pop1	66.021
WS-Raw2	44,392.297
WS-Pop2	54.670

http://stats.wikimedia.org/EN/TablesArticlesTotal.htm
All article counts as of November 2010

DE (German)

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	2935	2935	2935	2935	2935	2935
	Missing	0	0	0	0	0	0
Mean		35263.16	820.72	558.29	2.69	107.28	4.83
Median		28931.00	519.00	268.00	1.00	85.00	3.00
Mode		23615[a]	264	0	1	58[a]	1
Std. Deviation		23908.491	1173.813	1105.976	8.025	85.648	13.227
Range		379533	21524	20890	211	1351	556
Minimum		19067	146	0	0	38	1
Maximum		398600	21670	20890	211	1389	557
Sum		103497362	2408804	1638568	7900	314881	14173

Wiki Art Count	1,200,000
Population	132,100,000
Word weighting	0.856
Image weighting	50
Link weighting	5
Link weighting - External	20
WS-Raw1	13,785.825
WS-Pop1	104.359
WS-Raw2	8,507.685
WS-Pop2	64.403

FR - French

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	2849	2849	2849	2849	2849	2849
	Missing	0	0	0	0	0	0
Mean		43925.31	1066.98	628.29	5.01	146.69	5.18
Median		35083.00	625.00	189.00	3.00	112.00	3.00
Mode		25637	439	0	2	80	2
Std. Deviation		31072.314	1473.170	1409.974	12.336	122.776	7.161
Range		357448	24349	21362	229	1673	168
Minimum		19980	222	0	0	46	1
Maximum		377428	24571	21362	229	1719	169
Sum		125143220	3039835	1790005	14269	417921	14758

Statistics

Wiki Art Count	1,000,000
Population	177,600,000
Word weighting	0.932
Image weighting	50
Link weighting	5
Link weighting - External	20
WS-Raw1	15,697.135
WS-Pop1	88.385
WS-Raw2	9,398.627
WS-Pop2	52.920

IT - Italian

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	2981	2981	2981	2981	2981	2981
	Missing	0	0	0	0	0	0
Mean		45610.68	1009.75	609.75	6.80	167.53	4.60
Median		34431.00	582.00	191.00	3.00	113.00	3.00
Mode		24749[a]	339	0	1	78[a]	2
Std. Deviation		39674.869	1510.117	1469.043	21.778	164.218	8.298
Range		647673	20126	20299	503	3031	214
Minimum		18765	173	0	0	42	2
Maximum		666438	20299	20299	503	3073	216
Sum		135965430	3010065	1817655	20285	499420	13726

Wiki Art Count	758,000
Population	63,900,000
Word weighting	0.874
Image weighting	50
Link weighting	5
Link weighting - External	20
WS-Raw1	12,967.499
WS-Pop1	202.934
WS-Raw2	7,315.987
WS-Pop2	114.491

ES - Spanish

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	2952	2952	2952	2952	2952	2952
	Missing	0	0	0	0	0	0
Mean		37942.39	1047.79	714.57	3.24	120.50	6.16
Median		29642.00	540.50	243.50	1.00	89.00	4.00
Mode		20720[a]	301	0	1	85	2
Std. Deviation		31079.325	1729.532	1611.278	9.822	133.801	14.997
Range		595303	29361	29529	249	3912	612
Minimum		18386	168	0	0	34	1
Maximum		613689	29529	29529	249	3946	613
Sum		112005943	3093064	2109396	9565	355721	18179

Wiki Art Count	676,000
Population	415,333,333
Word weighting	0.932
Image weighting	50
Link weighting	5
Link weighting - External	20
WS-Raw1	9,672.214
WS-Pop1	23.288
WS-Raw2	6,431.845
WS-Pop2	15.486

CY - Welsh

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	3000	3000	3000	3000	3000	3000
	Missing	0	0	0	0	0	0
Mean		28759.90	416.28	191.70	1.53	110.59	2.62
Median		25597.50	323.00	95.00	1.00	90.00	2.00
Mode		20236	204	0	1	47[a]	2
Std. Deviation		12617.731	391.122	387.736	3.132	72.920	2.443
Range		203994	7467	7515	98	639	63
Minimum		15890	116	0	0	31	2
Maximum		219884	7583	7515	98	670	65
Sum		86279697	1248825	575102	4589	331781	7856

Wiki Art Count	30000
Population	695196
Word weighting	0.793
Image weighting	50
Link weighting	5
Link weighting - External	20
WS-Raw1	234.429
WS-Pop1	337.213
WS-Raw2	84.251
WS-Pop2	121.190

GA -
Irish

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	2913	2913	2913	2913	2913	2913
	Missing	0	0	0	0	0	0
Mean		32907.63	590.05	319.90	2.38	122.12	2.82
Median		28441.00	364.00	104.00	1.00	98.00	2.00
Mode		18544[a]	249[a]	0	1	52[a]	2
Std. Deviation		28216.033	854.483	772.239	8.114	166.445	1.964
Range		1248807	18138	12364	291	8147	48
Minimum		16355	8	0	0	0	0
Maximum		1265162	18146	12364	291	8147	48
Sum		95859932	1718802	931871	6947	355729	8202

Wiki Art Count	12000
Population	424823
Word weighting	1.009
Image weighting	50
Link weighting	5
Link weighting - External	20
WS-Raw1	126.314
WS-Pop1	297.332
WS-Raw2	59.800
WS-Pop2	140.765

GD - Gaelic

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	2942	2942	2942	2942	2942	2942
	Missing	0	0	0	0	0	0
Mean		29225.70	337.81	113.16	1.10	117.52	2.80
Median		28123.00	253.00	36.00	1.00	110.00	2.00
Mode		28206[a]	246	0	0	134	2
Std. Deviation		10661.349	260.886	239.284	2.106	69.059	2.031
Range		140042	3336	3233	40	900	23
Minimum		16424	120	0	0	32	2
Maximum		156466	3456	3233	40	932	25
Sum		85982019	993851	332928	3243	345753	8230

Wiki Art Count	8,100
Population	65,593
Word weighting	1.041
Image weighting	50
Link weighting	5
Link weighting - External	20
WS-Raw1	61.599
WS-Pop1	939.106
WS-Raw2	18.534
WS-Pop2	282.560

GV - Manx

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	2940	2940	2940	2940	2940	2940
	Missing	0	0	0	0	0	0
Mean		45402.27	561.83	275.29	2.84	194.30	3.49
Median		38633.50	401.00	138.00	2.00	166.50	2.00
Mode		33779[a]	339	0	2	216	2
Std. Deviation		29840.171	762.779	741.324	7.808	139.338	7.213
Range		384001	14518	14619	184	1715	168
Minimum		15360	104	0	0	30	1
Maximum		399361	14622	14619	184	1745	169
Sum		133482671	1651775	809340	8345	571252	10264

Wiki Art Count	3,600
Population	1,330
Word weighting	1.072
Image weighting	50
Link weighting	5
Link weighting - External	20
WS-Raw1	50.708
WS-Pop1	38,136.127
WS-Raw2	18.247
WS-Pop2	13,723.257

KW - Cornish

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	2863	2863	2863	2863	2863	2863
	Missing	0	0	0	0	0	0
Mean		34072.58	391.23	56.33	2.10	151.25	2.21
Median		31804.00	350.00	.00	2.00	136.00	2.00
Mode		34085	358	0	2	199	2
Std. Deviation		12208.881	207.293	157.937	2.001	80.566	.693
Range		168031	3520	3091	37	536	8
Minimum		16045	114	0	0	31	2
Maximum		184076	3634	3091	37	567	10
Sum		97549784	1120085	161286	6009	433043	6327

Wiki Art Count	2,000
Population	215
Word weighting	0.932
Image weighting	50
Link weighting	5
Link weighting - External	20
WS-Raw1	18.275
WS-Pop1	84,999.579
WS-Raw2	4.033
WS-Pop2	18,759.905

BR - Breton

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	2930	2930	2930	2930	2930	2930
	Missing	0	0	0	0	0	0
Mean		29030.88	423.48	152.23	2.03	107.02	2.71
Median		25680.50	300.00	28.00	1.00	85.00	2.00
Mode		23632	241	0	0	71	2
Std. Deviation		13246.867	500.186	478.508	3.971	73.810	2.641
Range		254037	12500	12114	129	1043	77
Minimum		16560	8	0	0	0	0
Maximum		270597	12508	12114	129	1043	77
Sum		85060479	1240809	446047	5959	313561	7943

Wiki Art Count	36,000
Population	371682
Word weighting	0.932
Image weighting	50
Link weighting	5
Link weighting - External	20
WS-Raw1	280.341
WS-Pop1	754.250
WS-Raw2	107.228
WS-Pop2	288.495

NL - Dutch

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	2945	2945	2945	2945	2945	2945
	Missing	0	0	0	0	0	0
Mean		32046.76	642.44	363.21	3.62	115.76	4.48
Median		26386.00	399.00	141.00	2.00	85.00	3.00
Mode		20214[a]	273	0	1	71[a]	2
Std. Deviation		35074.563	1070.417	862.033	10.133	164.490	20.427
Range		1434853	32747	16489	342	7111	1069
Minimum		17470	157	0	0	41	2
Maximum		1452323	32904	16489	342	7152	1071
Sum		94377714	1891980	1069665	10663	340909	13198

Wiki Art Count	656,000
Population	22,910,097
Word weighting	1.023
Image weighting	50
Link weighting	5
Link weighting - External	20
WS-Raw1	7,420.822
WS-Pop1	323.911
WS-Raw2	4,211.913
WS-Pop2	183.845

FY - West Frisian

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	2901	2901	2901	2901	2901	2901
	Missing	0	0	0	0	0	0
Mean		26798.30	425.42	205.33	1.87	96.77	2.74
Median		23330.00	328.00	94.00	1.00	71.00	2.00
Mode		16876[a]	326	0	0	39	2
Std. Deviation		13260.878	424.816	381.156	8.611	70.736	1.542
Range		300864	8257	8039	291	812	39
Minimum		14051	8	0	0	0	0
Maximum		314915	8265	8039	291	812	39
Sum		77741856	1234129	595666	5417	280740	7960

Wiki Art Count	18,000
Population	522,333
Word weighting	0.955
Image weighting	50
Link weighting	5
Link weighting - External	20
WS-Raw1	139.197
WS-Pop1	266.490
WS-Raw2	61.978
WS-Pop2	118.657

FRR - North Frisian

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	1959	1959	1959	1959	1959	1959
	Missing	0	0	0	0	0	0
Mean		28621.53	317.45	108.39	2.31	106.03	2.60
Median		25625.00	248.00	64.00	2.00	87.00	2.00
Mode		18277	239	0	3	82[a]	2
Std. Deviation		12643.373	203.482	181.084	1.776	62.192	1.618
Range		113474	2202	1977	20	424	23
Minimum		14962	99	0	0	27	1
Maximum		128436	2301	1977	20	451	24
Sum		56069582	621877	212344	4531	207722	5093

Wiki Art Count	1,000
Population	9,333
Word weighting	0.955
Image weighting	50
Link weighting	5
Link weighting - External	20
WS-Raw1	7.493
WS-Pop1	802.854
WS-Raw2	2.712
WS-Pop2	290.521

STQ - Saterland Frisian

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	2957	2957	2957	2957	2957	2957
	Missing	0	0	0	0	0	0
Mean		32041.55	390.90	155.78	2.03	123.92	2.80
Median		26188.00	285.00	56.00	2.00	101.00	2.00
Mode		24507	241	0	1[a]	55	2
Std. Deviation		18725.720	359.061	339.068	3.464	79.394	2.727
Range		181457	7131	7136	101	779	53
Minimum		8591	5	0	0	0	0
Maximum		190048	7136	7136	101	779	53
Sum		94746867	1155905	460648	5998	366444	8279

Wiki Art Count	2,000
Population	3,000
Word weighting	0.955
Image weighting	50
Link weighting	5
Link weighting - External	20
WS-Raw1	17.396
WS-Pop1	5,798.711
WS-Raw2	6.124
WS-Pop2	2,041.209

DA - Danish

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	2936	2936	2936	2936	2936	2936
	Missing	0	0	0	0	0	0
Mean		32106.82	624.81	335.21	2.52	117.80	4.25
Median		25603.50	365.00	93.00	1.00	88.00	3.00
Mode		1353	226[a]	0	1	63	2
Std. Deviation		31254.324	1047.931	988.403	6.309	107.178	7.253
Range		1226825	23000	23016	105	2286	162
Minimum		1353	16	0	0	0	0
Maximum		1228178	23016	23016	105	2286	162
Sum		94265616	1834452	984175	7406	345863	12489

Wiki Art Count	139,000
Population	5,650,000
Word weighting	0.955
Image weighting	50
Link weighting	5
Link weighting - External	20
WS-Raw1	1,438.980
WS-Pop1	254.687
WS-Raw2	738.519
WS-Pop2	130.711

IS - Icelandic

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	2929	2929	2929	2929	2929	2929
	Missing	0	0	0	0	0	0
Mean		28482.36	425.89	166.21	1.83	108.66	3.22
Median		24353.00	283.00	28.00	1.00	82.00	2.00
Mode		19050[a]	198	0	1	47	2
Std. Deviation		15773.563	489.186	456.041	5.353	101.198	3.534
Range		406014	6892	6848	171	2944	74
Minimum		11581	97	0	0	30	1
Maximum		417595	6989	6848	171	2974	75
Sum		83424830	1247422	486819	5359	318256	9430

Wiki Art Count	30000
Population	283333
Word weighting	0.896
Image weighting	50
Link weighting	5
Link weighting - External	20
WS-Raw1	235.126
WS-Pop1	829.856
WS-Raw2	91.458
WS-Pop2	322.792

CA - Catalan

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	2915	2915	2915	2915	2915	2915
	Missing	0	0	0	0	0	0
Mean		39474.18	838.49	448.92	3.91	138.78	3.74
Median		30613.00	516.00	144.00	1.00	95.00	2.00
Mode		24695	227	0	1	76	1
Std. Deviation		31342.948	1289.473	1165.352	14.020	135.857	7.013
Range		493906	30082	29320	386	1633	190
Minimum		18974	161	0	0	38	1
Maximum		512880	30243	29320	386	1671	191
Sum		115067245	2444196	1308614	11397	404538	10912

Wiki Art Count	296000
Population	10843387
Word weighting	0.905
Image weighting	50
Link weighting	5
Link weighting - External	20
WS-Raw1	3,835.679
WS-Pop1	353.734
WS-Raw2	2,003.373
WS-Pop2	184.755

GL - Galician

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	2944	2944	2944	2944	2944	2944
	Missing	0	0	0	0	0	0
Mean		35028.78	725.20	410.25	2.93	121.85	3.53
Median		28372.00	444.50	123.00	1.00	91.00	2.00
Mode		22353	337[a]	0	1	64[a]	2
Std. Deviation		25311.296	1068.713	1061.892	10.496	105.037	5.630
Range		565260	21276	21281	240	2308	154
Minimum		17446	5	0	0	0	0
Maximum		582706	21281	21281	240	2308	154
Sum		########	2134998	1207781	8635	358725	10388

Wiki Art Count	65000
Population	3395000
Word weighting	0.815
Image weighting	50
Link weighting	5
Link weighting - External	20
WS-Raw1	708.751
WS-Pop1	208.763
WS-Raw2	358.611
WS-Pop2	105.629

SC - Sardinian

Statistics

		num_bytes	num_wd_raw	num_wd_act	num_images	num_links	num_links_external
N	Valid	2250	2250	2250	2250	2250	2250
	Missing	0	0	0	0	0	0
Mean		28880.25	380.41	99.11	1.60	107.83	2.26
Median		26813.50	307.00	.00	1.00	95.00	2.00
Mode		19223[a]	377	0	2	139	2
Std. Deviation		9743.742	352.645	314.422	2.404	56.781	.972
Range		87530	6253	5605	49	431	22
Minimum		16052	116	0	0	28	1
Maximum		103582	6369	5605	49	459	23
Sum		64980563	855912	222995	3593	242620	5090

Wiki Art Count	2500
Population	1415000
Word weighting	0.874
Image weighting	50
Link weighting	5
Link weighting - External	20
WS-Raw1	17.640
WS-Pop1	12.467
WS-Raw2	5.292
WS-Pop2	3.740